M.Sc. Thesis
Master of Science in Engineering

DTU Compute
Department of Applied Mathematics and Computer Science

Unveiling Diversity in Wikipedia:

Analysis of Human Dimensions Across Pages and Contributors

Alejandra Navarro Castillo (s222712)

Kongens Lyngby 2024

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Matematiktorvet
Building 303B
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary

Digital technologies have dramatically transformed society, enhancing communication and access to information while also revealing and amplifying biases, particularly gender biases. Wikipedia, a major online reference, mirrors these disparities, with its contributor base predominantly male and limited female participation. This study investigates gender diversity within Wikipedia, focusing on both contributor demographics and content representation. By employing natural language processing, statistical analysis, and network theory, the research aims to uncover insights into gender influences on Wikipedia biographies and their linguistic characteristics. Text analysis showed that female contributors write fewer words in men's biographies and emphasize gender in women's articles and family and personal terms in general. Male editors associate women with family terms as well and use more abstract language that may emphasize undesirable attributes. Hyperlink network analysis revealed women's Wikipedia articles are generally more isolated than men's, hindering exploration of women personalities. Additionally, there's limited connectivity between male and female biography pages, with women's articles often linking to men's, but not vice versa. These findings highlight ongoing gender gaps on Wikipedia, stressing the need for balanced representation to ensure less biased portrayals.

Preface

This master thesis was prepared at the department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring a Master of Science and Engineering degree in Human-centered Artificial Intelligence.

Kongens Lyngby, August 4, 2024

Alejandra Navarro Castillo (s222712)

Acknowledgements

Very interesting topic, good technical knowledge, and great supervision and accompaniment have been the key points for me to finish my master thesis. When I started my master programme at DTU, I was lucky to find the courses that motivated me for the topic of my thesis.

These courses had both the same course responsible: Sune Lehmann. When it was time to select a topic, I was eager to apply my technical skills to a social issue, so I reached out to him.

I am very grateful to Sune for giving me the opportunity to collaborate with him and his research group. I would also like to thank Jonas Lybker Juul and Silvia De Sojo for their valuable advice, guidance, and support throughout this journey of research, exploration and learning.

And lastly, big thanks to Giulio for motivating me and reading through my entire thesis on the last day before submission.

Contents

List of Figures

List of Tables

Notation

In this thesis, the following conventions for numerical notation are used:

- **Comma (,)**: Used as a thousands separator in large numbers. For example, 10,203,202 denotes ten million, two hundred three thousand, two hundred two.
- **Period (.)**: Used as a decimal point to separate the integer part from the fractional part of a number. For example, 4.34 represents four and thirty-four hundredths.

CHAPTER 1

Introduction

Since the last decades, we have witnessed how the development of technologies has changed the world as we used to know it. This digitalization of our lives has modified many aspects of society and the interactions between humans. Digital technologies and online platforms are shaping how we interact, work, learn, and entertain ourselves [Hyn18]. It has, of course, brought us many clear advantages, such as fast communication and almost endless sources of information. However, these advancements have also come with unexpected (or unconsidered) downfalls.

As humans have developed strong technologies that everyone can use, they should be also responsible for diminishing the risks that these entail. It is therefore extremely important to realize, study and understand the limitations of information and communication technologies (such as biases, etc). These new platforms, far from being neutral, often reproduce and reinforce the biases and inequalities of the offline world (UN, n.d.). This is especially evident in the case of gender bias, which remains a persistent issue in online spaces, mirroring and sometimes exacerbating societal norms [Her08; OS18].

The gender gap is pervasive across many platforms [FKW18; Kas+20], and of course Wikipedia, the multilingual, free online encyclopedia that relies on the collective contributions of a volunteer community, is no exception [Bey+22; Cab+18; GLM15; Fer+21; Wag+15; Wag+16]. Wikipedia, being the forth most popular site in the world after Google, Youtube and Facebook [Sem24] and the largest and most-read reference work in history [Cha21], is a good reference to study and understand how these biases are present in today's technologies.

The Wikimedia Foundation, which hosts Wikipedia, has recognized the imbalance in its contributor base since at least 2010. A study conducted in collaboration with UNU-MERIT revealed that less than 15% of Wikipedia's editors were women [GSG10]. Since then, researchers, journalists, and Wikipedia community members have scrutinized these participation disparities. Various studies have found that only 8.5% to 16.1% of Wikipedia editors are women [Coh11; Min+21], while the participation of non-binary editors is nearly nonexistent [Ste17].

Moreover, discursive features of content such as language usage, hyperlink net-

work of articles, source availability, image representation, multilingual coverage, and article classification have also been proved to be gender biased, supposing an underdevelopment of women biographies [GLM15; Wag+15; Wag+16].

1.1 Scope

The fundamental goal of this study is to explore the gender diversity in Wikipedia across pages and contributors, with a focal point on the text content of the Wikipedia biographies. Whereas past studies have focused solely on the text bias of the articles, this master thesis will also explore how the gender of the editors also influences the way they portray the personalities in Wikipedia. Through natural language processing methods, statistical analysis and network theory, we wish unveil insight about this topic.

To achieve this, we will focus on the following research questions:

- Gender representation in Wikipedia contributors and Wikipedia biographies:
  - What is the gender distribution among Wikipedia editors?
  - What is the gender distribution among biography pages on Wikipedia?
- How does gender influence the representation and portrayal of topics in Wikipedia articles, in particular in biographies?
  - How gender contribute to variations in text length across different Wikipedia articles?
  - What common linguistic patterns and styles are evident in Wikipedia articles across different genders in biographies?
  - How do editor genders impact the linguistic characteristics of Wikipedia content?
  - How does the content of Wikipedia articles differ across various subject categories?
- What structure emerges from the hyperlink network of articles? Are there differences between biographical categories?

1.2 Outline

The rest of this thesis consists of 6 chapters, briefly presented here. In Chapter 2, an overview and evaluation of the relevant literature on the topic is presented, going through a general introduction about Wikipedia as a information and communication technology platform, passing through the gender bias in both participants and content

in Wikipedia and finally acknowledging the importance of language when it comes to gender inequality. In Chapter 3, I describe the provided and engineered datasets used for the later analyses. Chapter 4 is devoted to the methodology, tools and description of procedures for implemented in this study. In Chapter 5 the results from the multiple analyses are presented. And lastly, in Chapter 6 and 7 we present our findings, attempt to understand the results obtained, consider future research on the topic and provide a conclusion of the thesis.

CHAPTER 2

Literature review

2.1 Gender bias in digital spaces and in information and communication technologies

Information and communication technologies (ICTs) have become an integral part of modern societies, ubiquitously affecting daily life, from business and entertainment to health and education. Globally, between 2000 and 2023, Internet World Stats reports a dramatic increase in the number of internet users, rising by over 1,392% [Sta24].

Despite a significant increase in usage of the internet and other ICTs around the world, there exists a gender gap: women, especially in developing countries, tend to be on the wrong side of the digital divide. The digital divide refers to the gap between individuals, households, businesses, and geographic areas at different socio-economic levels regarding their opportunities to access ICTs and the internet. This access is measured by various scenarios: access to the usage of digital devices; skills necessary to effectively use digital technologies; availability, affordability, and quality of internet access; and variations in the ability to access relevant and beneficial online content and services. While there are still gender differences in access to ICT in developing countries, second-level digital divide[1] issues are more of a concern in developed countries [AS23].

There are several reasons for the gender digital divide, and they differ across countries and contexts. Some of these reasons are related to the gender inequality in economy and education, cultural expectations and traditional gender roles or policy failures to tackle systematic inequalities [Fou20]. Studies highlight the sociocultural factors as particularly significant. These sociocultural factors center on prevailing stereotypes that challenge and frequently diminish women's interest in and aptitude for technology, resulting in negatively affecting women's self-esteem and confidence in using technology [AS23; ED18; SK20; CFD17].

Research indicates significant differences in how men and women utilize technol-

[1] Second-level digital divide: The second-level digital divide, also known as the "usage divide" or "skills divide," refers to the gap between individuals who have access to digital technologies and the internet but differ in their ability to effectively use these technologies.

ogy, particularly the internet. Men predominantly engage in online activities such as information retrieval and entertainment. In contrast, women are more inclined towards using the internet for communication and social interaction [Jac+01; GD15; Dix+14]. These variations reflect broader gender disparities in communication styles and social behavior in digital environments, highlighting the nuanced ways in which different genders navigate and utilize online resources.

2.1.1 Wikipedia

Wikipedia, being a free, web-based, collaborative encyclopedia project allows users to edit and create articles on a wide array of topics, making it one of the largest and most frequently consulted reference works available online. The platform operates under a model of open collaboration, where anyone with internet access can contribute, though contributions are subject to review and modification by other editors. Wikipedia's extensive reach, with millions of articles in various languages, serves as a crucial resource for knowledge dissemination and a significant gateway to the broader web ecosystem. Studies reveals that English Wikipedia significantly contributes to external web traffic, generating 43 million clicks to other websites within a month [Pic+21]. Wikipedia often acts as a bridge between search engines and third-party websites, fulfilling information needs unmet by search engines alone. These findings highlight Wikipedia's role not only as an important information source but also as a significant gateway to the broader web ecosystem.

Being a powerful medium for disseminating information quickly and cost-effectively, Wikipedia has the potential to perpetuate social and cultural biases [FW17]. This concern is particularly significant given Wikipedia's influential role in shaping public understanding across various topics. Research indicates that the portrayal of role models on Wikipedia can significantly influence educational and career decisions, with tangible impacts on economic outcomes [Hin19]. Emphasizing the importance of representation, Beaman et al. [Bea+09] found that increased visibility of female leaders can undermine gender role stereotypes and reduce bias. Similarly, highlighting non-binary identities challenges cisnormative assumptions and disrupts traditional notions of gender [Rob22]. Moreover, these biases not only shape the perspectives of Wikipedia's vast readership but can also be amplified by computational models, given the platform's growing use as a data source [FW17]. This underscores the critical need for ongoing efforts to address and mitigate gender bias [Red+20].

2.2 Editors gender bias in Wikipedia

Wikipedia, being a public encyclopedia, has a big number of contributors. Moreover, anyone can edit Wikipedia and become a *Wikipedian*[2]. There are currently 47,472,432 Wikipedia accounts, but only a minority of users contribute regularly (120,065 have edited in the last 30 days[3]) [Wik24b].

The Wikipedia mission statement centers on the *empowerment and engagement of people around the world to collect and develop educational content* (see Wikimedia Foundation Mission Statement, accessed 30-05-2024 [Fou24]). While participation in Wikipedia is ostensibly open to all, one must consider whether this inclusivity is truly reflected in practice. Are there discernible patterns among those who contribute to Wikipedia?

Studies attribute online participation in the Wikipedia community to personal motivation, cultural and linguistic factors, and antecedents of participation [Oko+12]. In regards to personal motivation, research has revealed that Wikipedians express a strong desire to give back in an effort to enhance public knowledge of complex phenomenon, heightened intrinsic motivation, and heightened altruistic behaviors and motivations [BP10; CCC10; SH09].

Potential gender differences in many of these characteristics have not yet been identified on Wikipedia, which is likely due to the use of unstandardized measures or lack of exploration and/or reporting of gender in data analyses [Oko+12]. Prior studies indicate that women are more altruistic than men, specifically when the cost of the behavior is expensive [AV01]. This finding implies that women would be more active on Wikipedia. However, recent research has revealed a Wikipedia gender gap; women edit Wikipedia at significantly lower volumes than men [CB12; Lam+11; ES13; GSG10; HS13]. This suggests that even if women have a strong desire to contribute on Wikipedia, they may experience significant barriers when attempting to do so.

The 2011 Wikipedia editor survey [Wik11], which sampled 4930 editors from all language projects, found that 9% identified as female, 91% as male, and less than 1% as transsexual or transgender. Subsequent user surveys [Wik24a] have consistently shown that approximately 85% of the users identify as male. Alternatively, scholars have used publicly available data to estimate the gender distribution of contributors on the English Wikipedia. They found that the proportion of female editors was estimated to be around 16.1% in 2009 [Lam+11], 18% in 2011 [Ant+11], and 16% when analyzing data from 2001 to 2021 [BLS21]. However, these studies do not provide data on contributors to the English Wikipedia who identify as non-binary.

[2]Wikipedians are volunteers who contribute to Wikipedia by editing its pages, unlike readers who simply read the articles.

[3]Accessed: 30-05-2024.

Research has also uncovered disparities in editing activity, showing that fewer than 10% of Wikipedians are responsible for over 90% of the total contributions on the English Wikipedia. [OGR08]. In addition to this more general inequality, the identification of a gender disparity, in which men edit at greater volumes on the English Wikipedia, was revealed in recent Wikipedia surveys [GSG10; HS13; Pan11], analyses of Wikipedia itself [Lam+11], and subsequent research about the media's often belittling responses to reports about the Wikipedia gender gap [ES13].

Explanations for this gender gap often highlight the contentious nature of Wikipedia as favoring traditionally male communicative styles [CB12; Lam+11; Lan+12], as women may be less sure of their expertise, more often targeted for harassment, and more negatively affected by critical feedback than men [BC16]. Consequently, women may contribute more during online discussions rather than through visible, article-based Wikipedia editing [Lam+11]. Gender differences in Internet familiarity [HS15], the desire to contribute to the common good, and differences in leisure time may also contribute to the gender gap.

2.3 Biography pages' content gender bias in Wikipedia

Wikipedia, as one of the largest and most popular sources of information, reflects societal biases that extend into its content, particularly in the representation of gender in biography pages. Numerous research studies have consistently shown that the biographies of women are significantly underrepresented compared to men, both in terms of the number of articles and the coverage quality. Also, the interactive project *Humaniki* highlights this prevalent imbalance [Hum24] and as in May 2024, out of the 4.45 million humans with at least one Wikipedia article, the percentage of *women with at least one Wikipedia article* is 18.830% compared to men 81.084% (note that the percentage of *other genders* is very low, only 0.086%).

However, this numbers have been even more extreme in the past. Whereas today the percentage of women biographies in the English Wikipedia is 19.8%, six years ago, in 2018, the data showed that only 17.7% of English Wikipedia biographies were about women [WZ18]. Despite a 1.8% increase over five years, the persistent gender gap underscores the ongoing challenge in achieving gender parity on the platform [Qai+22].

The content of the article biographies on Wikipedia also reflects gender bias. The platform's biography pages exhibit a notable bias in the topics and professions covered, with significant implications for gender representation [Gua14].

A predominant bias in Wikipedia's biography pages is evident in the over-representation of biographies related to traditionally male-dominated professions and fields. For instance, categories such as American football players, tech entrepreneurs, and political leaders predominantly feature men. Conversely, professions and topics traditionally associated with women, such as care-giving roles, education, and arts, are often underrepresented or less prominently featured.

Another core principle on Wikipedia is *notability*. Wikipedia defines this term as a test used by editors to decide whether a given topic warrants its own article in the encyclopedia. To be deemed notable, a topic must have received significant coverage in reliable, independent secondary sources. These sources must provide evidence that the topic is of enduring interest and importance, transcending trivial or transient events. This concept plays a critical role in shaping the content and scope of Wikipedia, influencing which topics, individuals, and subjects are considered worthy of inclusion.

Despite the assumed consensus among Wikipedians, the supposedly "neutral" roles and formalities on the site often involve subjectivity and bias in their application and impact [Luy12]. Studies show that women's biographies are slightly more notable than men's [Wag+16], and the level of activity and traffic on Wikipedia articles dedicated to female scholars are not proportionate to their scientometric achievements [SY14]. Scholars have repeatedly voiced concern that wiki-notability is inconsistently enforced, arbitrarily assessed, and biased against women [GS17; LR09; Vit17].

Despite the high notability of female biographies, Tripodi [Tri23] found that there is significant gender bias in the deletion nominations of Wikipedia biographies. Over a 38-month period, despite women comprising less than 19% of all biographies, they represented over 25% of those nominated for deletion each month. Furthermore, women's biographies were more frequently miscategorized as non-notable, with roughly 25% of them being incorrectly labeled and later retained, compared to only 17% for men. These findings create an additional challenge in addressing Wikipedia's content gender gap, as the survival of these biographies relies on editors increasingly participating in contentious, male-dominated deletion discussions, rendering the current system unsustainable.

However, gender bias in Wikipedia content extends beyond article selection, manifesting in various textual and visual elements. Multiple studies have explored these imbalances, revealing how gender is depicted in encyclopedic content [BW22]. For instance, empirical evaluations have found that women's articles tend to be longer than articles about men, although this may be influenced by the asymmetry in article coverage [GLM15; Wag+15]. Additionally, gender-specific word associations and a greater emphasis on social relationships and family in women's biographies highlight further biases in the text [Wag+16]. Visual bias is evident in the uneven use of images observed by [Bey+22; ZFW17]. Finally, the use of references in content construction

has shown a source asymmetry, with women's biographies having more references and diverse sources compared to men's [YWK16]. These discursive disparities contribute to the overall content gender gap observed in Wikipedia.

2.4 Gender bias in language

The relationship between language and thought has been a subject of intense debate and exploration among linguists, philosophers, and cognitive scientists for centuries. This intricate connection raises fundamental questions about how language influences our perceptions, shapes our reality, and structures our cognitive processes [Bor06].

There is great evidence to think that language actively shapes and constrains our thinking patterns (the Sapir-Whorf hypothesis or linguistic relativity hypothesis) [Bor11; Fra23]. According to this view, the structure, vocabulary, and concepts present in a language shape the cognitive processes and the worldview of its speakers. Also, it was found that language plays a crucial role in how human conceptualize and categorize the world around. It provides the framework for organizing and labeling experiences, objects, events, and abstract concepts. The specific categories and distinctions made in a language can influence how individuals perceive and interpret their environment.

The field of study that explores how language interacts with society is called *sociolinguistics*. It examines the ways in which language varies and changes in different social contexts, and how these variations and changes reflect and influence social factors including cultural norms, expectations, and context. The study of language variation is concerned with social constraints determining language in its contextual environment. The variation can be associated with various societal aspects like: socioeconomic status, age, gender, ethnicity, level of education, etc [CP98].

The study of language and gender has produced a body of diverse results for gender differentiation in language use. Research indicates that men often use numbers, technology-related terms and URLs [BES14; Sch+06; Ngu+13], while women use more family and relationship-oriented language [BO05]. Grammatical structure features have also been examined, revealing that men typically use more prepositions and articles, whereas women use more pronouns [Arg+03; Sch+06; NO06; Arg+07; Arg+09; Ott10; Sch+13; BES14]. Stylistic features show that men tend to use longer words [Sin01; GSR09; Ott10], while women employ more emotional language [NO06; Sch+13; BES14].

Wagner et al., in their study on Wikipedia articles [Wag+15], found that articles about women tend to emphasize the fact that they are about a women (i.e., they contain words like *woman, female* or *lady*), while articles about men do not contain

words like *man, masculine* or *gentleman*. The reduced prominence of male-related terms in articles about men may be linked to the idea of male as the "null gender" [FJR06], indicating a social bias that perceives male as the default gender in certain social contexts.

Assuming the importance of language and how influential it is in the way humans think, the study of how the different editors in Wikipedia write their revisions is of great importance to shape the gender bias in this online encyclopedia.

2.5 Text analysis with natural language processing tools

With the intention of exploring how social factors affect language and variation, the sociolinguists have to go through different stages of the research like data collection, data analysis, interpretation of results and reflect on the findings.

Data collection is a fundamental step in the research cycle for researchers in both sociolinguistics and computational linguistics. Traditionally data collection involved manual, labor-intensive methods like observation, surveys, and interviews, resulting in smaller data sets. Inevitably, these data collection methods are labor-intensive and time-consuming and the resulting data sets are often small. However, the rise of computer-mediated communication and social media has revolutionized data collection, providing large, easily accessible sources of both informal and formal language [And17]. This marks a significant shift from manual to computational methods. The integration of computational techniques to handle large-scale data sets represents a major methodological change [Ngu+16].

When it comes to methods for researching about language and text, originally, linguistics combined symbolic methods and linguistic theory [KR96]. Over time, with the advent of large corpora and large-scale statistical methods and modeling frameworks, there was a shift away from symbolic and knowledge-driven methods towards more empirical, data-driven approaches [BP22].

This is where natural language processing (NLP) comes in. NLP can be defined as the field of research that studies how computers can understand and process data encoded in human (natural) language, with the goal of completing tasks like speech recognition, text classification, natural-language understanding, and natural-language generation, among others. It is therefore an interdisciplinary subfield of computer science, artificial intelligence and linguistics [AS18; BKL09].

NLP techniques can help with various aspects of text processing and analysis, going from the initial steps like data collection, extraction and cleaning of the text,

or text processing (tokenization, stemming, converting to lowercase, removing stopwords, etc) to the analysis itself through statistical methods and machine learning models [BKL09; AS18].

The ensemble of techniques in NLP enables quantifying the information contained in text by automated text analysis, allowing researchers to not only track the presence or prevalence of particular terms and ideas, but also to measure relationships between them, and how those relationships change over time. In short, natural language processing provides a powerful tool to help understand people and culture.

2.6 Improving the gender gap in Wikipedia

Concerns about gender imbalances on Wikipedia emerged in 2010, with the Wikimedia Foundation, led by Sue Gardner, aiming to increase female contributors to 25% by 2015. However, by 2014, co-founder Jimmy Wales admitted this goal had not been met. Academic attention to gender issues on Wikipedia began in 2007, with initial research on motivation factors and contributions. The first paper focused on gender and Wikipedia appeared in 2009, and significant research on the topic emerged in 2011, quantifying the gender gap in contributions and topics [Lam+11] and activity levels [Ant+11]. Despite projections of gender parity by 2034 [KK15], this did not consider the influence of women in article deletions [Tri23]. Studies found women's involvement in the Wikipedia community to be low, with policies often opaque and enforced by experienced editors furthering their own agendas [CB12; Mor+13; Jem20]. The feminist perspective highlighted the gender gap further in 2016, explored in depth by Ford and Wajcman [FW17].

Women Wikipedians create safe spaces within and outside Wikipedia due to their unique experiences with the gender gap and safety issues. Simple interventions, such as proactive moderation and positive feedback, can increase female contributions. Smaller Wikipedia communities are perceived as more welcoming for women, although the link between the gender gap and harassment needs further study [Fer+21; MEP19].

Feminist interventions, like women-only edit-a-thons, combat gender inequality by creating secure editing environments. These events, where women create and edit feminist content and address misogynistic language, have become popular and effective in narrowing the gender gap. They have increased the coverage of African women and provided educational experiences that empower participants to challenge biases on Wikipedia [BC16; Ukw+21]. There are also other initiatives like *Women in Red (WiR)* [Wik24c], which is a group of editors committed to improve systemic bias on Wikipedia and close the gender gap by focusing on creating content regarding women's biographies, women's work, and women's issues. Their name derives from the practice of turning "red links" (pages that do not yet exist on Wikipedia) into

blue (an active page).

Studies also indicate that female mentorship is crucial for encouraging the inclusion of more women in Wikipedia. Network analysis shows that female mentorship promotes women's inclusion across various fields. It has been seen as crucial to pay attention to the portrayal of women on Wikipedia and adopt a more gender-balanced vocabulary when writing articles [KK21; LB19].

Efforts to attract more women to Wikipedia are yielding positive results, but there is a need for greater gender parity among influential editors who shape policies and perform high-level tasks. Addressing the gender disparity among the most active editors and adopting an inclusive approach in platform design and governance are essential for promoting equality on Wikipedia [MEP19; Ant+11; Lam+11].

CHAPTER 3

Dataset

This chapter will outline the dataset utilized in this project, beginning with the data collection process and including a description of the transformations applied to clean and pre-process certain elements of the collected data. Finally, it will address the ethical considerations related to this dataset.

3.1 Data collection

The dataset used in this study was created by extending, combining and refining some initial data that was made available by the Department of Applied Mathematics and Computer Science, Technical University of Denmark, in order to conduct relevant research as part of this Master Thesis project.

The initial dataset was collected and combined for a previous research study [Tor23] on the same topic and was stored in the DTU Compute servers. The underlying dataset is a Wikipedia dump, provided by the Wikimedia Foundation, containing metadata for all revisions that modify any Wikipedia namespace made to English Wikipedia from January 2001 to February 2023. Additional datasets were incorporated to filter the data and to include additional variables to compose this initial dataset.

3.1.1 Initial dataset considerations

The dataset was filtered to include metadata from Wikipedia pages and Wikipedia talk pages only. Anonymous users and bots were excluded, reducing the dataset significantly. The dataset was further filtered to include only biography pages with corresponding gender information. Using the Wikipedia Human Gender Indicators dataset and the Wikidata Query Service, biography pages without gender information were removed, reducing the dataset to 110,183,743 observations. Professional categories for the biographies were extracted from the DBpedia Ontology, which uses information from Wikipedia infoboxes. Each biography was associated with one category, though some pages defaulted to the parent category *Person*.

After applying all filters, the final dataset includes metadata for revisions made to biography pages and talk pages in English Wikipedia from January 2001 to February 2023. The dataset comprises 36,995,319 rows representing edits and 226,988 rows representing comments.

This data provides a valuable foundation for my current investigation, enabling me to build upon previous findings and explore new dimensions of the topic at hand.

3.2 Data cleaning and pre-processing

To gather the required data for addressing the research questions, the initial dataset was filtered to include only the most recent revisions of all Wikipedia pages. Additionally, it was modified to encompass the complete text of the Wikipedia pages of the filtered revisions and the gender of all editors who contributed to each specific revision.

3.2.1 Filtering the dataset for the last revision

For the purpose of addressing the research questions, I aimed to analyze the most recent data available in the initial dataset. Therefore I filtered the initial dataset to include only the most recent revision of each Wikipedia page. Through this filtering, the dataset got reduced to 1,281,779 observations (unique Wikipedia pages).

In the original dataset, each row corresponds to a revision made by an editor, containing comprehensive information about the revision (e.g., revision ID, timestamp) as well as details about the editor (e.g., editor ID, username, gender). For the purpose of my analysis, I focused exclusively on the most recent revision, identified by the newest timestamp. This necessitated considering all editors who contributed up to that specific revision. Consequently, alongside processing the information pertaining to the last revision, I also compiled and analyzed data regarding the group of editors who made contributions.

Once this filtering was made, the total number of unique editors is 128,876 and the gender distribution is ‘Male’ 85.44%, ‘Female’ 14.42% and ‘Other’ 0.13%. This numbers align with the aforementioned studies about the gender of the editors (see Section 2.2 of the literature review).

3.2.2 Extracting the wikitext and wikilinks

My research is mostly focused on the language utilized by the contributors when they write about people in the Wikipedia pages, so, in order to complete my dataset, I

needed the text of the specific articles.

The method to retrieve the content of a specific Wikipedia revision was using the Wikipedia API [Med24]. Initially, I defined the URL for the API endpoint and set the necessary parameters for the API request, including the action ("query"), the format ("json"), and properties to retrieve revisions by their ID and include the content. After getting the response from the API request, the response is converted into JSON format, so that it can be processed by Python, and the wikitext content of the specified revision is obtained by navigating through the JSON structure. This allows for the extraction and further analysis of the revision's full text.

For this study, also the wikilinks were extracted. The wikilinks refer to internal links to Wikipedia pages that are included in the texts of other Wikipedia pages. These links point to other articles within Wikipedia, helping to create a web of interconnected content. The list of wikilinks for each specific revision of the dataset was retrieved again by the Wikipedia API. Setting the URL for the Wikipedia API endpoint and defining the parameters for the API request, specifying the revision ID and the properties to get, which are "text" and "links". Then, the JSON response from the API was parsed and a list of links is derived.

3.2.3 Cleaning the wikitext

The extracted content has the format of *Wikitext*[1]. Wikitext, also known as wiki markup or wiki code, is the syntax and markup language used by Wikipedia and other MediaWiki-based websites. It allows users to create and edit web pages easily without needing extensive knowledge of HTML or other web programming languages. An example of a wikitext obtained from one of the Wikipedia pages used in this study is presented in appendix A.1.

The wikitext encompasses the complete content of the article, including all headings, paragraphs, lists, tables, and other elements as they appear in the specified revision, all formatted in wikitext markup.

With the aim of analysing the written language of editors, the wikitext needs to be pre-processed into written natural language. I undertook several steps to transform the wikitext into readable text, enabling further text analysis. The method is systematically processing a given wikitext by applying a series of pattern-based transformations through regular expressions[2] designed to remove unwanted elements and simplify the text. The text processing procedure involves the following steps:

1. Removing the outermost `{{ ... }}` blocks.

[1] https://en.wikipedia.org/wiki/Help:Wikitext

[2] Regular expressions (often abbreviated as regex or regexp) are sequences of characters that form search patterns as form of strings.

2. Removing HTML comments within `<!--` ... `-->` tags and HTML tags within `<` ... `/>`.
3. Removing reference tags `<ref>` and `</ref>`.
4. Processing the links retaining only the relevant text within `[[link]]` and `[[link|text]]` formats.
5. Removing file links in the `[[File: ... ]]` format.
6. Simplifying section titles within = characters.
7. Removing category tags starting with `Category:`
8. Deleting hyperlinks starting with `[https:`

By applying these transformations using regular expressions, I produced a clean and more readable version of the text, becoming natural written language text, for every Wikipedia article in the dataset.

3.2.4 Data overview

After all the process of gathering, filtering and cleaning the data, the final dataset is composed by 1,281,779 pages divided into 24 page categories.

3.3 Ethical considerations

The original dataset was composed from different sources: a readily available and public Wikipedia dump and publicly available gender information associated with the Wikipedia contributors.

Under the european General Data Protection Regulation (GDPR [Con18]), usernames are regarded as personal data because they can identify an individual, either directly or indirectly. Similarly, gender is classified as personal data since it can be linked to an identifiable individual, either on its own or when combined with other data [Koc20]. However, according to the GDPR, the personal data collected in this instance is considered non-sensitive, meaning no specific data processing measures are required, nor is express consent necessary (Art. 6, GDPR). Therefore, I am invoking the legitimate interest of the researcher to operate with this data for the purpose of conducting the designated study.

The completion of the original dataset with the extraction of the wikitext and links of each of the articles was carried through the Wikimedia API, which gives open public access to all Wikimedia content.

Only the necessary data for the research purposes outlined in this project has been used. The data has been securely stored on DTU Compute servers, following the university's recommendations. Any presented results or disclosed information are in aggregated and anonymous form. Lastly, the personal data will be deleted once it is no longer needed.

CHAPTER 4

Methods

The overall goal of this work is to examine potential gender inequalities in the content of Wikipedia articles (specifically biographies) along different dimensions. Focusing on the relation between the contributors, (also referred to as editors) who identify as male, female, and non-binary, and the Wikipedia biographies (also referred as pages or articles), the study analyzes their contributions on the text of the articles across different dimensions: length of the articles, lexical level[1], and structural level (links).

4.1 Initial exploration

The initial analysis offers a broad overview of gender representation across the Wikipedia pages from different page categories[2] both for editors and biographies. This stage of the study addresses the following research questions:

- What is the gender distribution among Wikipedia editors participating in Wikipedia biographies? How does this distribution varies depending on the biography category?
- What is the gender distribution among biography pages on Wikipedia? How does this distribution varies depending on the biography category?

To address these research questions, the study conducts a descriptive analysis. This involves visual methods and tables presenting the frequencies and percentages of relevant variables.

As we are considering only the last revision of every Wikipedia page, many contributors may have participated in the modification of this newest revision. Therefore, looking at the gender of the group of editors for the pages that are analysed in this study is of big importance. Two ways of measuring the gender of the group of editors were considered: "majority" and "totality".

[1]The lexical level of language concerns words and their meanings, including their use and relationships.

[2]This study define *page categories* (also referred as *biography categories, biographical categories* or *article categories*) as the professional field associated to each Wikipedia page.

- **"Majority" measure**: all the editors and their gender are considered and the gender of the group is just the gender that appears more times.

- **"Totality" measure**: all the editors and their gender are considered and the gender of the group is only considered if all the editors of the group have the same gender. Otherwise, the gender of the group is considered to be "No" (there is no totality of one gender).

4.2 Differences in text content

In this study, a focal point is the investigation of the language used in the Wikipedia articles. As mentioned previously in sections 2.3 and 2.4, linguistic bias is a systematic asymmetry in language patterns as a function of the social group of the persons described, and is often subtle and therefore unnoticed. Also, how language is used can be an indicator of the characteristics of the writer, and particularly, their gender and their biases. Therefore, the research questions examined in this main part of the study are the following:

- Are there any text content differences between biographies of different gender? Are there differences in the text content between editors of different gender? In terms of:
  - Length of texts,
  - Lexical analysis: for the study of the semantics and linguistic style,
  - and Part-of-speech tags: for the analysis of discourse and linguistic style.
- If there are differences, how significant are they across biographies' gender and contributors' gender? And what does this indicate?

This part of the study is focused on finding differences between the text content in articles about women and men, and edited by female and male editors. Therefore, the way that the gender of the group of editors were measured was the "totality" measure explained in the previous section. This means that only the pages that were edited by contributors with the same gender are considered in this part of the analysis. Thus, the four study groups considered are the following:

	Group of female editors	Group of male editors
Female page	Group 1	Group 2
Male page	Group 3	Group 4

Table 4.1: Four study groups

4.2.1 Classification of texts

A preliminary method to uncover differences between groups is to apply Machine Learning (ML) classification models. The goal of this analysis is to determine whether these models can effectively differentiate between the genders of contributors and the genders of biographies. If the models demonstrate the ability to classify correctly, surpassing the performance of a baseline model, it would indicate that there are notable differences in the text that merit further investigation. With this in mind, it is important to note that optimizing model performance is not the focus of part; therefore, no parameter tuning or cross-validation is performed when training the models.

For this classification experiment, only the pages from the *Person* category were employed, since this category contains the most diverse (due to their nature of different topics) and largest amount of pages. I will denote the texts of this subset of articles as corpus[3] C.

Two approaches were considered:

- Classification of the gender of the group of editors (2 classes: Male and Female)
- Classification of the gender of the pages and gender of the group of editors (4 classes: the ones defined in table 4.1).

An important thing to take into account for both of the classifications is the unbalanced class representation of the dataset. For the category *Person*, the number of pages for each group are displayed in table 4.2. Due to this reason, random undersampling was performed to balance all classes to the same number of articles as the smallest class.

	Group of female editors	Group of male editors
Female page	5,449	58,696
Male page	4,369	183,254
Total	**9,818**	**241,950**

Table 4.2: Number of articles divided into four study groups for the *Person* category.

Next, when working with textual data, text processing is a crucial step in transforming raw text data into a structured numerical format suitable for analysis and machine learning. Consequently, the texts (also called *documents*) from corpus C were preprocessed to obtain a set of features that the models could interpret. The td-idf (Term Frequency-Inverse Document Frequency) representation was used to transform the text data into numerical feature vectors. This technique is commonly used in natural language processing (NLP) and machine learning to convert text data

[3]In linguistics and natural language processing (NLP), a *corpus* (plural: *corpora*) refers to a large and structured set of text or spoken data that is used for analysis and research purposes.

into a format that can be used by machine learning algorithms.

The tf-idf measure is composed of two terms:

- **Term Frequency (tf)**: The number of times a term appears in a document. This measures the importance of a term within a specific document.

$$tf(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d} \tag{4.1}$$

$$= \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{4.2}$$

- **Inverse Document Frequency (idf)**: A measure of how important a term is within the entire corpus. It helps to reduce the weight of terms that appear very frequently across documents and increase the weight of terms that appear less frequently.

$$idf(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing term } t}\right) \tag{4.3}$$

$$\approx \log\left(\frac{1+n}{1+|\{d : d \in D \text{ and } t \in d\}|}\right) + 1 \tag{4.4}$$

Tf-idf: The product of tf and idf, providing a weight that indicates the importance of a term in a document relative to the entire corpus.

$$tfidf(t,d) = tf(t,d) \cdot idf(t) \tag{4.5}$$

It is to note that there are different variants of how to measure the inverse document frequency (idf) weight. The formula for $idf(t)$ in equation 4.4 adds the constant 1 to the numerator and denominator of the idf as if an extra document was seen containing every term in the collection exactly once, which prevents zero divisions. Additionally, the effect of adding 1 to the idf in the equation is that terms with zero idf, i.e., terms that occur in all documents in a training set, will not be entirely ignored.

In order to calculate the tf-idf score of the words in the corpus, the `TfidfVectorizer`[4] class from the library `sklearn` (Scikit-learn) from Python was utilized. This method converts a collection of raw documents to a matrix of tf-idf features by first converting the collection of text documents to a matrix of token counts and then transforming the count matrix to a normalized tf-idf representation.

To achieve this objective, several steps of text preprocessing needed to be performed for each article in the corpus:

[4] `TfidfVectorizer` is a class in the `sklearn.feature_extraction.text` module of the Scikit-learn library from Python. Documentation here: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

	born	university	woman	work	film	art	new	also	life	national	award
0	0.000000	0.000000	0.00000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000
1	0.000000	0.000000	0.00000	0.000000	0.000000	0.0	0.000000	0.032727	0.000000	0.0	0.000000
2	0.026298	0.000000	0.00000	0.000000	0.079896	0.0	0.000000	0.015931	0.017729	0.0	0.000000
3	0.096629	0.000000	0.00000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000
4	0.023587	0.016675	0.00000	0.016153	0.000000	0.0	0.018043	0.042868	0.031804	0.0	0.021607
...	...	...	...	...	...	...	...	...	...	...	...

Figure 4.1: Example of tf-idf representation of documents. Each document of the corpus is a row in the matrix and the columns represent the words (or terms) of the corpus. The values in the matrix are the tf-idf(t, d) of each document d and term t.

1. **Remove punctuation**: this was achieved by utilizing the `re` library in Python, which is a module that provides support for working with regular expressions.
2. **Convert to lowercase**: by using `.lower()` method in Python, that converts all the characters in a string to lowercase.
3. **Tokenize the text**: process of breaking down a text into smaller units, which are called tokens. In this case, the units were words. This was performed using the `word_tokenize` module from the `nltk` library in Python.
4. **Remove stopwords**: stopwords are common words in a language that are considered to have little or no significant meaning. These stopwords were obtained from the `nltk` library and some other words were added.

Finally, after these steps, all documents of the corpus have a numerical representation which consists of the tf-idf value of each term that appears in each document. A subset of the representation from corpus C is showed in figure 4.1.

Once the data is processed to the desired format for the ML models, the classification was performed and evaluated accordingly. A fundamental step in machine learning is the train-test split, which is used to evaluate how well a model generalizes to new, unseen data. Therefore, the dataset was randomly divided into two separate subsets: one for training the model (75% of the dataset) and one for testing it (the rest 25%).

For the classification, several models where used to account for variability in model performance, feature handling, computational efficiency and interpretability. The following models were used:

- Baseline model: predicts the majority class for all instances.
- Logistic Regression.
- k-Nearest-Neighbours (kNN) with $k = 1$.

- Support vector machine.

It is important to remember that in this part of the study, the classification models are just a motivation to explore for differences in the texts of the different groups of articles based on gender of the page and editors. Consequently, the ML models parameters have not been optimized, since this was not the aim of part of the analysis.

After having trained the models, we have to see how well they performed. The measure selected for evaluation of the classification is the accuracy of the prediction, because it provides a quick and simple measure of how well a model is performing. The accuracy is calculated by measuring the proportion of correctly classified instances out of the total instances.

$$\text{accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

4.2.2 Number of words in text

As we want to measure as many text characteristics as possible to try to see the differences in page gender and editors gender, I started by analyzing the length of the texts by the number of words. The analysis was conducted comparing the four study groups mentioned in table 4.1 across every page category. Specially, I wanted to compare the way that female and male groups of editors differ.

It is important to note that some of the categories analyzed had very small sample sizes, particularly within the groups of female editors, meaning there were very few articles edited by female contributors. For example, when examining the *Cleric* category, despite it being relatively large, the contributions by female editors were minimal: only 6 groups of female contributors edited women's biographies and 19 groups edited men's biographies.

The analysis employed in this study utilizes straightforward statistical techniques, along with confidence intervals that are derived from bootstrap methods [BS12]. This approach allows for a robust assessment of the data by generating multiple resamples, thereby providing a more comprehensive understanding of the variability and uncertainty associated with the estimates. Through this combination I aim to enhance the reliability of the findings.

4.2.2.1 Bootstrap approach for hypothesis testing and confidence intervals

The bootstrap sampling approach was chosen for its versatility, particularly when dealing with small sample sizes, as previously discussed was the case for some of the

page categories, and when the normality of the sampling distribution[5] cannot be assumed [ET94]. This technique is also ideal for easily calculating the standard error of the statistic being studied.

The bootstrap method is a resampling technique used to estimate statistics on a population by sampling with replacement from the original sample data. What we want to compare is if the differences between female and male groups of editors are significant, i.e. if the data coming from these two groups follow the same distribution or not.

$$\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n_F} \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_M} \end{pmatrix}$$

The procedure that has been performed is the following: We have the two genders of groups of contributors (*female* and *male*), and we want to test if there is a statistically significant difference in the median of number of words of the articles that these groups have edited. So the null hypothesis is H_0: the number of words is the same in both groups of contributors.

1. **The test-statistic chosen is the median of the number-of-words distribution.** The observed values for the statistic are calculated in the two groups: $\tilde{X}$ and $\tilde{Y}$ (*female* and *male* editors respectively).

2. **We determine the distribution of the test-statistic.** Bootstrap resamples ($B = 1000$) from each group are generated. For both groups, each resample has the same size ($n =$ size of the smallest group (normally, the female editors group)). Subsequently, the median is calculated for each bootstrap resample. We now have the two sampling distributions of medians for each group: $\left(\tilde{X}_1^*, ..., \tilde{X}_{1000}^*\right)$ and $\left(\tilde{Y}_1^*, ..., \tilde{Y}_{1000}^*\right)$. As a consequence of the Central Limit Theorem, the distribution of the bootstrap sample statistic will tend to be normally distributed as the number of bootstrap samples becomes large, even if the original data is not normally distributed.

3. **We estimate the confidence intervals for the statistic of each of the two groups.** The means ($\bar{X}^*$ and $\bar{Y}^*$) and the standard errors ($SE_{\bar{X}^*}$ and $SE_{\bar{Y}^*}$) of the sampling distributions are calculated. Having produced B bootstrap replicates $\tilde{X}_b^*$ of an estimator $\tilde{X}$, the bootstrap standard error is the standard deviation of the bootstrap replicates:

$$SE_{\bar{X}^*} = \sqrt{\frac{\sum_{b=1}^{B} \left(\tilde{X_b}^* - \bar{X}^*\right)^2}{(B-1)}}$$

[5]The sampling distribution of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample.

Same procedure for calculating $SE_{\tilde{Y}^*}$. Now, the confidence intervals for the median for a confidence level of $100(1-\alpha)\%$ can be obtained by:

$$\bar{X}^* \pm (SE_{\tilde{X}^*} \times z_{\alpha/2})$$
$$\bar{Y}^* \pm (SE_{\tilde{Y}^*} \times z_{\alpha/2})$$

where $z_{\alpha/2}$ is the standard normal value with probability $\alpha/2$ to the right. For this case, we have established $\alpha = 0.05$, so $z_{\alpha/2} = 1.96$.

4. **We interpret the results.** To assess the significance of the difference between the two groups of editors, we examine whether their confidence intervals overlap. If there is an overlap, no conclusions can be drawn. However, if there is no overlap, we can infer that the estimated medians are sufficiently distinct, suggesting they are unlikely to have resulted from random sampling variability alone. Consequently, we can reject the null hypothesis and conclude that the number of words in texts differs between the two groups of contributors.

4.2.3 Td-idf scores

Another key point when studying text characteristics is the usage of certain key words and the frequency of words in general (as part of a lexical analysis). The aim of this part of the study is to test whether some words have more importance in a certain document compared to other documents.

To analyse this aspect, I used term frequency–inverse document frequency measure (tf-idf), which is a measure of importance of a word to a document in a collection or corpus, adjusted for the fact that some words appear more frequently in general. This popular statistical technique became a fundamental tool in the fields of information retrieval and natural language processing, after the concept of idf was introduced by Karen Spärck Jones in a 1972 paper titled "A Statistical Interpretation of Term Specificity and Its Application in Retrieval" [Spa72]. This measure has already been used to process the text data for the classification models previously described in section 4.2.1 and the equations 4.2, 4.4 and 4.5 explain this measure in detail.

This exploration aims to compare differences in tf-idf scores for words in the documents edited by groups of female contributors versus groups of male contributors in each page category. For that reason, the corpus of texts were divided into categories and each of them were further divided into the women biographies and men biographies.

In an effort to answer the research question about "how significant are the differences in the importance of words used in documents edited by different contributors' genders", I used the same statistical method that was used to examine the number of words (see 4.2.2.1). In this case, the initial sample is the documents edited by each of

the groups of editors (i.e. female or male editors) within a category and article-gender. The studied test-statistic is the tf-idf score of a word.

$$\mathbf{X} = \begin{pmatrix} \text{Female edited document 1} \\ \text{Female edited document 2} \\ \vdots \\ \text{Female edited document } n_F \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} \text{Male edited document 1} \\ \text{Male edited document 2} \\ \vdots \\ \text{Male edited document } n_M \end{pmatrix}$$

With the intention of comparing the tf-idf scores of a word between the two groups of editors, for each bootstrap resample the selected documents were concatenated into a single text (obtaining a single string). This is done because of how the idf score is calculated, i.e. taking into account all the documents to compare and calculating the inverse document frequency. For more detail, see the infographic in figure 4.2. The approach to calculate the tf-idf score of a word in the corpus was to follow the same steps as in part 4.2.1: preprocessing the text and utilizing the `TfidfVectorizer` class from scikit-learn. Finally, the result of this process is a tf-idf representation of the two documents considered (female-edited and male-edited), containing the tf-idf scores for all the words in the documents of each bootstrap resample.

The aim of this method was to assess a diverse array of words to yield more comprehensive insights into the vocabulary utilized. The selection of specific words for this study was made thoughtfully, grounded in previous literature, ensuring that my analysis reflects established findings and contributes to a deeper understanding of language usage.

The Finkbeiner test [Fin13] suggests that articles about women often emphasize the fact that she is a woman, mention her husband and his job, her kids and childcare arrangements, how she nurtures her underlings, how she was taken aback by the competitiveness in her field and how she is such a role model for other women. Also, the historian Gillian Thomas, who investigated the role of women in Britannica, states in her book [Tho92] that as contributors, women were relegated to matters of "social and purely feminine affairs" and as subjects, women were often little more than addenda to male biographies (e.g., Marie Curie as *the wife of* Pierre Curie).

Based on these findings, I created the following four topics of words that capture some of the aspects that could be over-represented in articles about women:

- Gender topic contains words that emphasize that someone is a man or a woman: *women, woman, man, men, mr, mrs, lady, gentleman.*
- Sexuality topic consists of words about the sexual orientation: *gay, lesbian, heterosexual, homosexual.*
- Relationship topic aggregates words about romantic relationships: *married, divorced, couple, husband, wife, spouse.*

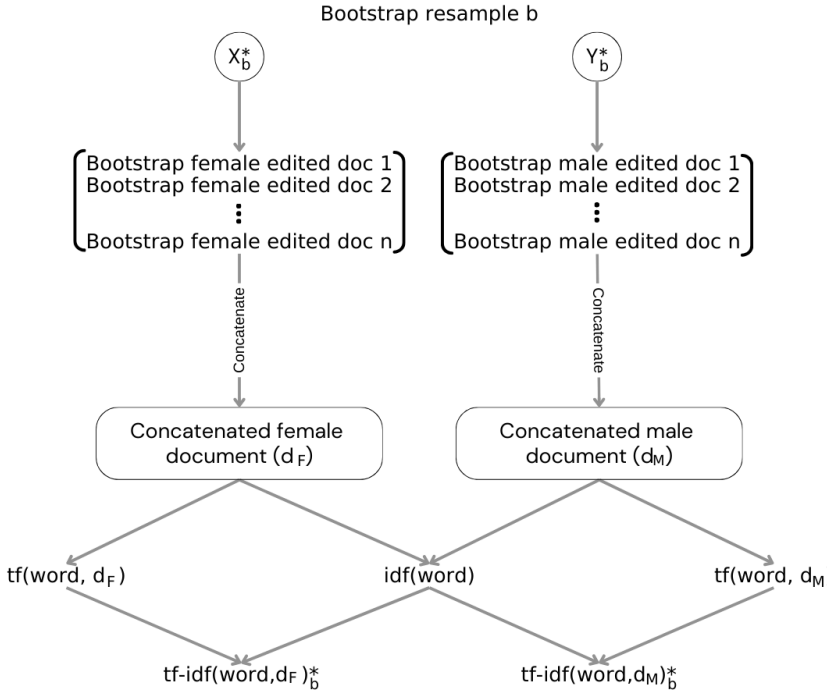


Figure 4.2: Process of calculating the tf-idf score of a word in a document for one bootstrap resample of the initial collection of documents: X (pages edited by groups of female contributors) and Y (pages edited by groups of male contributors).

- Family topic contains words about family relations: *kids, children, mother, father, grandmother, grandfather.*

4.2.4 Bigrams of words

In the previous section, we dug into the idea and approach to analyse single words within texts, examining their frequencies and significance. This exploration should provide insights into the most commonly occurring terms and their relative importance within the corpus. However, language is inherently structured beyond isolated words; it unfolds through sequences and patterns that span across consecutive terms.

In this part, I extend the analytical lens to encompass collocations[6], which capture

[6]Collocations are pairs or groups of words that are commonly used together in a language. These word combinations sound natural to native speakers and are often habitual pairings. For example, in English, we say "make a decision" rather than "do a decision" and "heavy rain" rather than "strong rain."

these sequential relationships. Collocations are used in language learning, lexicography, and linguistic studies to describe how words typically co-occur in speech and writing. The way to analyze collocations in computational linguistics, text mining, and natural language processing is through n-grams, which are contiguous sequences of n items (usually words or characters) from a given sample of text.

N-grams, ranging from bigrams (2-grams) to higher-order sequences, offer a more nuanced perspective on textual content by revealing collocations, phrases, and syntactic structures that uncover deeper layers of semantic cohesion and contextual relevance within the text. Building upon the understanding gained from single-word analysis, this section explores specifically the methodologies for the study of bigrams of words (pairs of consecutive words).

For the goal of identifying word combinations in a corpus that exhibit some idiosyncrasy in their linguistic distribution, this task typically —but not exclusively— involves comparing the statistical distribution of the combination to the distribution of its individual constituents using an association measure.

In statistics, probability theory and information theory, **pointwise mutual information (PMI)** [CH90] or point mutual information, is a measure used to determine the strength of association between two terms in a corpus. It measures the extent to which the actual probability of two events co-occurring, $p(x, y)$, deviates from the expected probability based on the individual probabilities of the events and the assumption of independence, $p(x)p(y)$. The concept was introduced in 1961 by Robert Fano [FH61] under the name of "mutual information".

The PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the probability of their coincidence given their joint distribution and their individual distributions, assuming independence. Mathematically:

$$\mathrm{PMI}(x, y) = \log_2 \left(\frac{p(x, y)}{p(x) \cdot p(y)} \right) \tag{4.6}$$

The measure is symmetric, meaning $\mathrm{PMI}(x, y) = \mathrm{PMI}(y, x)$. It can take positive or negative values, but is zero if x and y are independent. Positive values of PMI indicate that the two words co-occur more frequently than expected by chance. Higher positive values imply a stronger association. Negative values of PMI suggest that the two words co-occur less frequently than expected. This can indicate that the presence of one word makes the other less likely to appear.

In order to calculate the PMI of a pair of words in the texts, these need to be processed into the desired format. In this part of the analysis, each of the articles in the *Person* category were processed in the same way, described below:

1. Remove punctuation, remove stopwords, tokenize and convert the text to lowercase. These steps were performed as previously explained in 4.2.1.

2. Compute word frequencies: this was done using the `Counter` class from the `collections` module in Python. The `Counter` class is a specialized dictionary subclass used for counting hashable objects.

3. Obtain the bigram list: The `bigrams` function from `nltk.util` is then used to generate a list of bigrams from the tokenized words. Each bigram is a tuple containing two consecutive words from the text.

4. Compute the bigram frequencies: also executed with the `Counter` class from the `collections` module in Python.

Having extracted the frequency distribution of single terms and also bigrams, the way that the marginal probability of term x is calculated is as follows:

$$P(X = x) = \frac{\text{Count of term } x}{\text{Total number of words}} \tag{4.7}$$

and the joint probability of the bigram (x, y) is:

$$P(X = x, Y = y) = \frac{\text{Count of bigram } (x, y)}{\text{Total number of bigrams}} \tag{4.8}$$

For this part of the study, a set of specific bigrams has been chosen for analysis. When examining biographies of women, the bigrams take the form of *woman* + word, and similarly in men's biographies as *man* + word. The words selected for analysis include: *first, rich, black, white, old, young, family, married, children.*

This selection is due to several reasons: words like *first* can denote a position or rank; *young, old* reflect age-related descriptors; *rich, family, married* reveal societal norms or stereotypes related to wealth, family structure and marital status; *black, white* can be used as racial or ethnic descriptors.

In order to directly compare the gender of the editors using these bigrams, I used the bootstrap approach for confidence intervals described in 4.2.2.1, using as initial samples the articles edited by the two editors' genders (X and Y). Also in this case, for each bootstrap resample the selected documents were concatenated into a single text, in order to calculate the count of terms and bigrams for the entire ensemble of articles considered (similar procedure that the one displayed in figure 4.2).

Through this systematic exploration of bigrams, I aim to explore the context in which *woman* and *man* are described or discussed, shedding light on societal perceptions, roles, and attributes associated with gender in Wikipedia.

4.2.5 POS-tagging (or adjective/verb ratio)

Aside from lexical bias, previous research also indicates that linguistic biases in general may emerge when people describe individuals who belong to either their in-group or out-group[7] [Ott15].

The Linguistic Intergroup Bias (LIB) theory [Maa+89] suggests that for members of our in-group, we tend to describe positive actions and attributes using more abstract language, and their undesirable behaviors and attributes more concretely. In other words, we generalize their success but not their failures. Note that verbs are usually used to make more concrete statements (e.g., "he failed in this play"), while adjectives are often used in abstract statement (e.g., "he is a bad actor").

Conversely, when an out-group individual does or is something desirable, we tend to describe them with more concrete language (we do not generalize their success), whereas their undesirable attributes are encoded more abstractly (we generalize them). Maass et al. point out that LIB may serve as a device that signals to others both our status with respect to an in- or out-group, as well as our expectations for their behavior and attributes [Maa+89]. Our expectations are of course not only determined by our group-membership but also by the society in which we live. For example, in some situations or domains not only men but also women may expect other women to be inferior to men.

To investigate the question of how subjective the discourse is from both editors groups in their in- and out-groups, the approach will consists of comparing the overview of biographies about men and women in terms of their adjective/verb ratio.

The procedure followed for each article text was:

1. **Lowercase the text and tokenize it:** first by using `.lower()` method in Python, that converts all the characters in a string to lowercase. Second, by using `word_tokenize` module from nltk.

2. **Part-of-Speech tagging:** it is a process in natural language processing where words in a text are tagged with their corresponding parts of speech. This process involves assigning a part of speech label, such as noun, verb, adjective, etc., to each word in a sentence. This was performed using `pos_tag` module from nltk library.

3. **Counting adjectives and verbs.**

[7] In group: the group of people with whom an individual identifies as being a member. Out-group: the opposite of the in-group, those who an individual does not identify with or perceive as part of their own group.

4. **Calculating the ratio:**

$$r = \frac{\text{Number of adjectives}}{\text{Number of verbs}}$$

For more accuracy on comparing the adjective-to-verb ratio between genders in pages and contributors, the bootstrap approach for confidence intervals described in 4.2.2.1 was used. The initial samples were the articles edited by the female editors X and by the male editors Y. As in previous parts of the study, also now each bootstrap resample involved concatenating the selected documents into a single text to count the adjectives and verbs across the entire set of selected articles (similar procedure that the one showed in figure 4.2).

4.3 Hyperlinks structure

In this study, another way to explore diversity in gender in the Wikipedia pages is to study the articles' hyperlinks structure. One can describe this as the gender-specific tendencies to preferentially link articles about people to those of the same or different gender. For instance, one might hypothesize that articles about women contain more links to men or vice versa.

In this section, I wanted to explore the hyperlink structure of articles in Wikipedia, and the following research questions were investigated:

- What structure emerges from the hyperlink network of articles? and, are there differences between page categories?
  - Are there trends in the gender of isolated pages?
  - What is the degree of homophily?
  - Are there connectivity tendencies across biography genders?

In order to investigate the hyperlink network of articles, for every biographical category a network have been created. The nodes of the network are articles of the category, and these nodes are linked if the name of an article of the category appears in the text of the origin article. Therefore, it consists of a directed graph. Note that not all the pages in the category are included in the network, since only the biographies that contained a hyperlink to the any biography in the same category are considered.

Some measures have been investigated with the aim of exploring the gender patters in connectivity of Wikipedia articles.

4.3.1 Isolation

The isolated nodes in a network are a characteristic of the network that represent the number of nodes that have no connections or edges linking them to other nodes within the network. In the Wikipedia context, having hyperlinks in other pages is something of great importance when it comes to relevance and retrieval of articles [KK08; FWD19].

The measure utilized in this study is the following:

$$\text{Isolation ratio}(g) = I_g = \frac{\text{Number of isolated nodes of gender } g}{\text{Number of nodes of gender } g} \tag{4.9}$$

which provides a normalized measure of how prevalent isolated nodes are within the subset of nodes of gender g. A higher isolation ratio indicates a greater proportion of isolated nodes within that gender group.

In order to check for statistical differences in isolated nodes ratios between genders, I have used the **permutation test** [Wag+15; FC19]. The main idea of a permutation test is to determine the significance of an observed effect by comparing it to the distribution of effects obtained by randomly rearranging the data. Specifically, it involves repeatedly shuffling the data labels and recalculating the test statistic for each permutation to build a distribution of the test statistic under the null hypothesis. The null hypothesis of a permutation test is that there is no effect or no difference between the groups being compared ($H_0 : F = G$) where F and G are the distributions of the two groups being compared.

In the case of our hyperlinks network, the hypothesis that I wanted to investigate is whether the isolation ratio differs significantly in each gender page group. Therefore, the null hypothesis of this test is:

$$H_0 : \text{the isolation ratio is the same for female and male biographies.}$$

To perform the test, the chosen test statistic is the absolute difference of the isolation ratio between female and male nodes: $|I_F - I_M|$. Then, for each permutation p of the test, the gender attribute of each node are randomly shuffled and the test statistic is calculated $|I_F - I_M|_p$. A total of 1000 permutations were computed in this analysis. The p-value of the hypothesis is estimated as the proportion of permutations that give a difference as large or larger than the absolute difference of isolation ratio of the original sample. With a significance level[8] established at $\alpha = 0.05$, we will reject the null hypothesis when the p-value is greater than α.

[8]The significance level of a test is the probability of rejecting the null hypothesis when it is actually true, also known as the Type I error rate.

4.3.2 Homophily

Homophily is the tendency of individuals to associate and bond with similar others. To answer some of the research questions, studying the homophily of the articles network will give an idea on how much the nodes attach to others that are similar in some way. In this study, the gender of the page is the attribute that has been analyzed.

4.3.2.1 Assortativity coefficient

One way to measure homophily in a social network is the assortativity coefficient. It measures the tendency of nodes in a network to connect to other nodes that are similar in some way. In the context of social networks, this often means that individuals with similar attributes (e.g., age, education level, income) are more likely to be connected. The assortativity coefficient was measured for the attribute "gender of a page" with the standard definition of assortativity [New03]:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{4.10}$$

where, e_{ii} is the fraction of edges connecting nodes with attribute i to nodes with attribute i. a_i is the fraction of edges with source nodes having attribute i. b_i is the fraction of edges with target nodes having attribute i.

The assortativity coefficient ranges from -1 to $+1$, where -1 indicates perfect disassortativity, meaning that nodes tend to connect to other nodes that are different in the attribute being measured; 0 means no assortativity, meaning there is no preference for similarity or dissimilarity in the connections; and $+1$ signify perfect assortativity, meaning nodes tend to connect to other nodes that are similar in the attribute being measured.

4.3.2.2 Asymmetry

With the assortativity coefficient previously described, only the nodes of same gender are being considered. However, it is also of great interest to analyze the degree of asymmetry from one gender to the other that the links of Wikipedia articles have.

To assess asymmetry across genders, we need to calculate the probability that an article about a person of a certain gender links to an article about a person of the opposite gender. The L score measure is introduced in the 2015 paper by Wagner et al. [Wag+15] and it measures the log-likelihood ratio between edge probabilities. It is defined as follows:

$$L(g_1, g_2) = \log\left(\frac{P(\text{to} = g_2 \mid \text{from} = g_1)}{P(\text{to} = g_2)}\right) \tag{4.11}$$

where $P(\text{to} = g_2 \mid \text{from} = g_1)$ is the conditional distribution that an edge links to an article of gender g_2 given that it comes from an article of gender g_1, and $P(\text{to} = g_2)$ is the probability that any link ends in an article of gender g_2 regardless of the gender of its origin. We compare the probability that a link ends in an article of gender g_2 given that it comes from an article of gender g_1 with the probability that a link ends in an article of gender g_2 regardless of the gender of its origin.

This way, positive values of L indicate increased connectivity from g_1 to g_2, and negative values the opposite. This defines an assortativity matrix of the four combinations of genders that measures the tendencies to connect within and across genders.

For assessing the asymmetry between genders, we can examine the L score entries for one gender to the other. Wagner et al. defines a measure as follows:

$$A = L(F, M) - L(M, F) \tag{4.12}$$

where F correspond to the female pages and M to the male pages. Positive values of A will indicate a stronger tendency of women biographies to connect to articles about men than the opposite, controlling for the difference in in-degrees and sizes of both genders.

CHAPTER 5

Results

5.1 Gender and biographical category distribution

As an initial approach to the data, analyzing the number of articles per biographical category provides an overview of their distribution, revealing crucial information for further analysis. Figure 5.1 illustrates the distribution of pages across the different categories. The category with the highest quantity of articles is *Person*[1]. The disparities in article counts among categories will significantly influence several subsequent parts of the study.

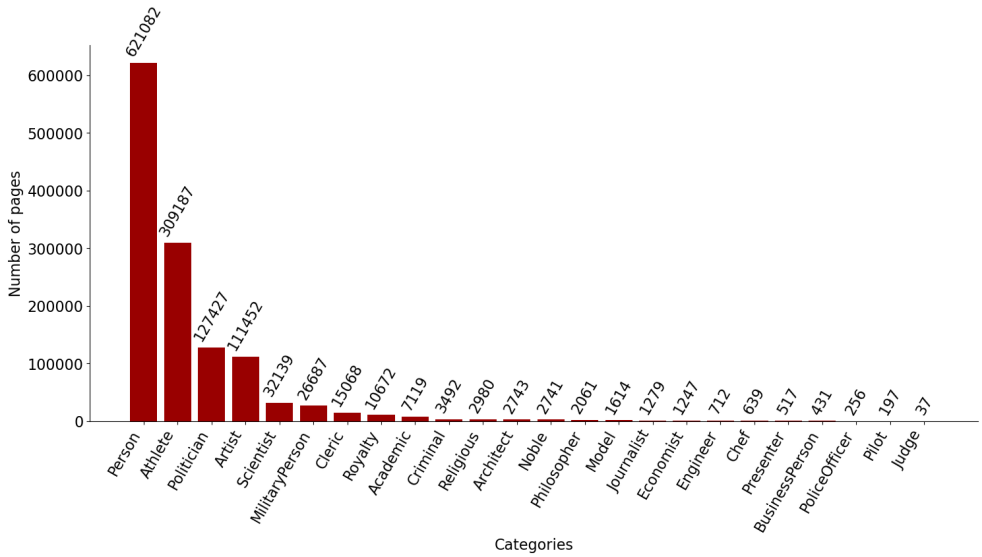


Figure 5.1: Number of pages (biographies) per category, sorted in descending order.

[1]The *Person* category contains all biographies that are not classified into a professional subcategory.

5.1.1 Gender of editors

In analyzing the gender diversity among editors contributing to Wikipedia biographies, the findings reveal a marked imbalance. The majority of contributors are male, representing 85.45% of the total (equating to 110,118 individuals). Conversely, females represent a significantly smaller portion of the editors at 14.42% (18,586 individuals). The representation of non-binary individuals is minimal, accounting for just 0.13% (172 contributors).

Figure 5.2 depicts the gender distribution among Wikipedia contributors for the different biography categories. Inspecting the graph, we observe that all categories are approximately equally under-representative of female contributors. All categories have a percentage of female editors lying between 10% and 20%.

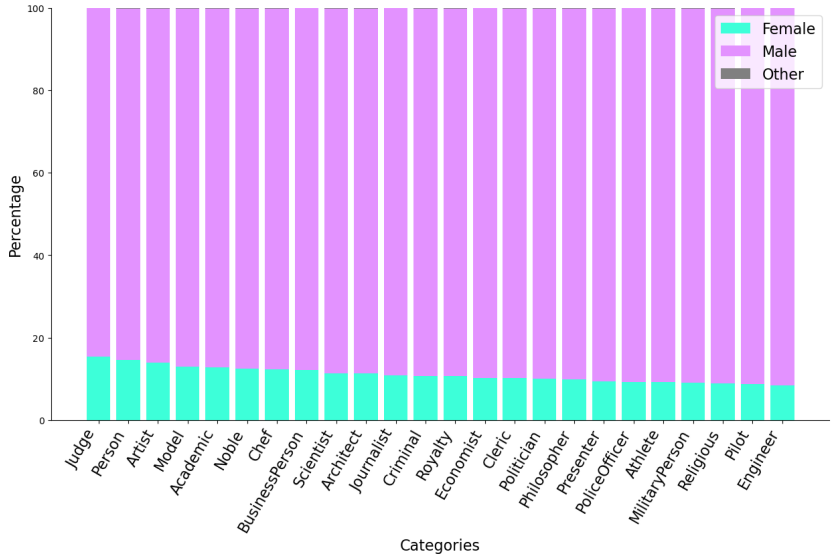


Figure 5.2: Editors' gender distribution across the different page categories sorted in descending order by percentage of female editors.

Given that only the latest revision of each Wikipedia page is considered, numerous contributors might have been involved in editing the most recent version. Two ways of measuring the gender of the group of editors were considered: "majority" and "totality" (explained in section 4.1). The analysis in figure B.1 shows the distribution of the gender of the group of editors measured with these two measures.

Looking at the results, when analyzing the "majority" of gender in plot B.1(a), one can see that again most of the groups of editors have majority of male contribu-

tors. This is however not surprising given the proportions of contributors' gender in Wikipedia. When examining the "totality" measure in B.1(b), it is noticeable that we find a remarkable amount of pages that have been edited only by editors of the same gender.

5.1.2 Gender of biographies

To complete the overview of gender imbalance, it is also important to see the gender distribution of the biographies of our study. As analysed in previous studies (see literature review section 2.3), also in the dataset analysed in this study there is a big imbalance in the gender of the biographies. The majority of biographies are male, comprising 74.63% of the total, which equals 956,536 pages. In contrast, females make up a much smaller portion of the biographies, representing 25.35% or 324,971 pages. Non-binary biographies are minimally represented, constituting only 0.02%, or 272 pages.

Regarding the distribution of gender across categories, the results are showed in figure 5.3. Most categories show that female representation does not exceed 40%, except for the *Model* category, which has over 85% female pages.

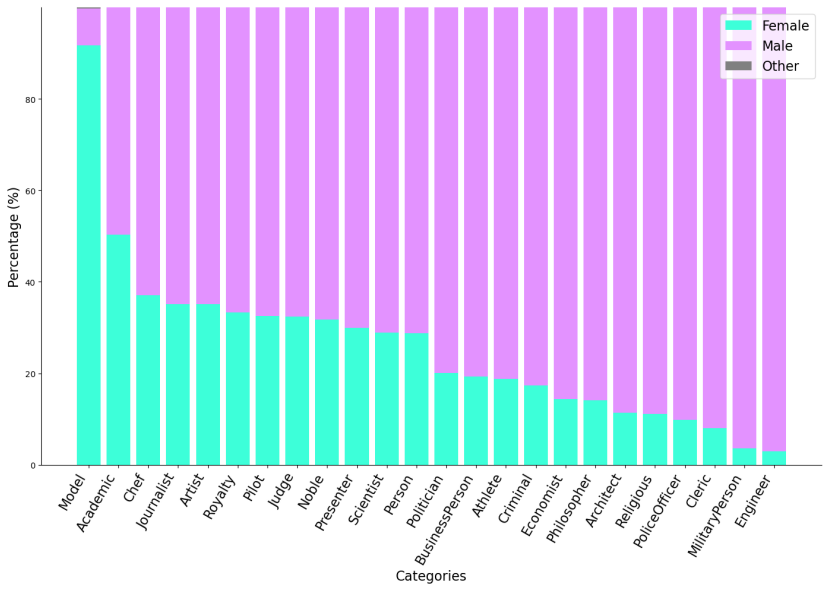


Figure 5.3: Distribution of biographies' gender across the different categories sorted in descending order by percentage of female biographies.

5.2 Differences in text content

In the previous section, we have seen that there is an imbalance in the gender of both the contributors and the biographies. But, when it comes to gender gap in Wikipedia, there are other factors to take into account. In this section, the differences in various characteristics of the texts edited by different gender groups were studied.

5.2.1 Classification of texts

Can ML models identify any differences in language when comparing between the gender of editors and pages in Wikipedia? The following table shows the results for the classification models used in the two classification tasks (defined in 4.2.1).

Model	**Accuracy**
Baseline	0.500
Logistic reg.	0.793
kNN	0.747

Table 5.1: Accuracy of the models when classifying 2 classes (only gender of the group of editors).

When inspecting the results of the editors' gender classification (table 5.1), it is visible that all the classification models seem to perform better that the baseline model, since their accuracy it is well above 50%. This indicates that even considering Wikipedia articles from both genders, the models can identify differences in the writing style of the two contributors' genders considered.

Model	**Accuracy**
Baseline	0.250
Logistic reg.	0.685
kNN	0.598
SVM	0.725

Table 5.2: Accuracy of the models when classifying 4 classes (gender of page and gender of group of editor combined).

Now, when analysing the 4-class classification results (table 5.2), on a first inspection we observe that the accuracy of the models seem to be much higher than the baseline model. This indicates apparent differences between texts of different classes. Then, is it possible to know what is causing these differences?

With a further inspection of the Logistic regression model, it is possible to check which coefficients have higher weights into the classification. Table C.1 in the appendix displays the 10 most positive and negative coefficients and shows to what

features (terms) correspond for the 4 classes classification.

As seen in the table, the word *woman* clearly differentiates the classes since it has the largest positive coefficient in female pages as well as the largest negative coefficient in male pages. There are also other words that have big impact in classifying the female biographies, and that is terms like: *husband, female, mother, daughter, child, girl, lady.* Most of these terms emphasize the gender of the female biographies by utilizing gender words. If we have a look at the male biographies, the terms whose coefficients are larger are *refer, john, son, election, footballer, james, charles* which indicates that there are probably a lot of biographies of people with those proper nouns. However, inspecting the male biographies, there are no signs of gender words, but on the contrary, there are more career specific words as *election, painter, footballer, chairman, director*, etc.

5.2.2 Number of words in texts

One significant aspect of content analysis in textual data is the length of the texts. This was meticulously measured by determining the word count in each individual text. The analysis was conducted separately on female and male biographies, as well as independently for each biographical category, to study the differences between editors' gender. The editors' gender was considered to be the "totality" measure (explained in section 4.1).

As a first result, we show in figure 5.4 a histogram representing the distribution of the number of words per document in the category *Person*. This distribution seems to follow a power-law distribution (note that the y-axis is represented in logarithmic scale) because of its heavy tail and fast decay. This is not surprising since the power-law distribution is a popular distribution for human-made phenomena [Sta78].

From figure 5.4, this time comparing the medians of the distributions across different editors' groups, a slight difference is noticeable: groups of female editors (labeled as "F_editors") seem to have a lower median in their distributions. This observation prompts the follow-up question: does this pattern hold true across all categories, and is the observed difference significant or merely due to chance?

The previous questions are explored with the help of the statistical test using bootstrap sampling explained in part 4.2.2.1 for each category. In terms of the statistical differences about the count of words in texts, results are displayed in figures 5.5(a) and 5.5(b), for both female and male pages respectively.

When examining women's biographies, the results indicate that while significant differences in word counts may exist within certain categories (like *Athlete*, *BusinessPerson*, *Model*, *Philosopher* and *Religious*), overall, there is no strong evidence to confirm substantial differences in text length when comparing female editors to

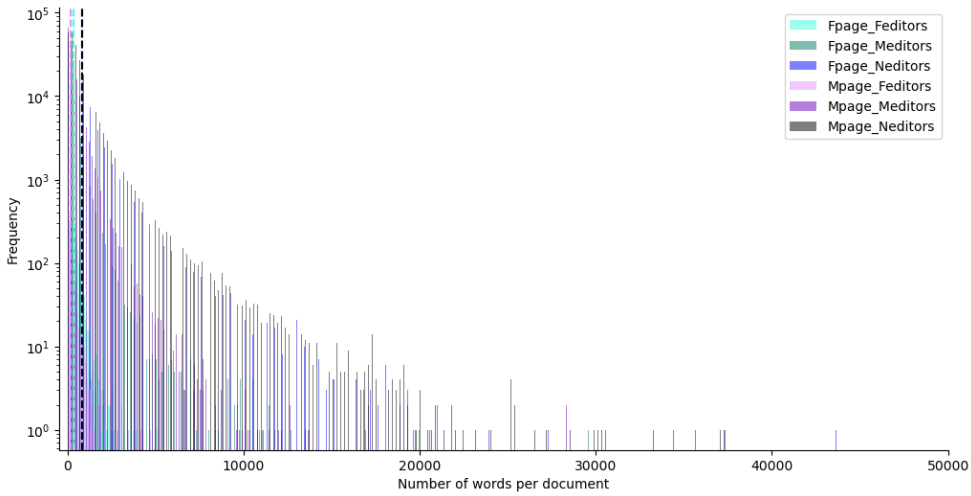
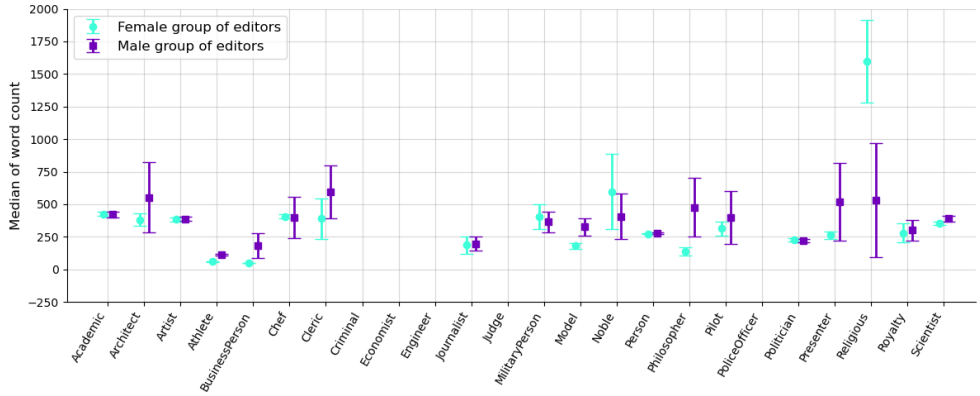


Figure 5.4: Histogram of the number of words per document in the *Person* category. The y-axis is in logarithmic scale. The dashed lines represent the median values for each distribution within the groups. These groups are categorized based on the gender (F for female, M for male, and N for "No" totality) for either biographies ("page") or editor groups ("editors").

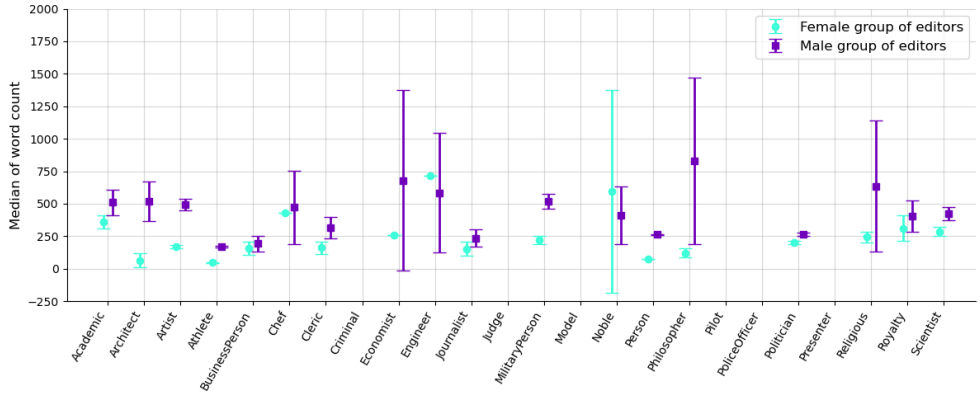
male editors across both female and male biographies. Also, in table C.2(a) in the appendix, looking at the confidence intervals and comparing between female and male contributors in the four largest categories (which have more significant results because of a larger sample), we see that the confidence intervals overlap (except in *Athlete* category), indicating no significant difference.

However, during the examination of the male biographies, there are some statistically significant differences between female and male editors. In most of the categories, it can be observed that the confidence intervals for the median of the sampling distributions from female editors and male editors do not overlap, so that it can be concluded that the estimated medians are sufficiently far apart that they are unlikely to have arisen from random sampling variability alone. This means that the median word count for men edited pages is significantly higher than for pages edited by women.

Table C.2(b) in the appendix displays the values for the confidence intervals of the median word count in men's biographies across the four largest categories. It is apparent that female contributors write significantly fewer words than their male counterparts, since these are the most meaningful categories due to their large amount of documents.



((a)) Female pages.



((b)) Male pages.

Figure 5.5: Confidence intervals of the median statistic for the word count of female and male editors pages, for each biographical category. The dot represents the mean of the sampling distribution of the median, while the error bar the lower and upper bound of the confidence interval. The absence of data in some categories is due to these pages not being edited by any female group of editors.

5.2.3 Tf-idf scores

The importance of a selection of words, described in 4.2.3, was measured in this part of the study.

As a first glance of the results, we will inspect the results of the category *Person*, since it is a general one and contains the largest amount of articles. In figure 5.6

the 100 words with the highest tf-idf score for each of the 4 groups (female/male biographies and groups of female/male editors) are showed in the form of word clouds. More specific information about the tf-idf scores of the 10 most important words can be found in table C.4 in the appendix.

((a)) Female biographies by female editors.

((b)) Female biographies by male editors.

((c)) Male biographies by female editors.

((d)) Male biographies by male editors.

Figure 5.6: Word clouds of each of the 4 groups (female/male biographies and groups of female/male editors) in the *Person* category.

It is quite notable that the word *women* has great importance in the female biographies (with tf-idf scores of 0.2605 and 0.1741 in female editors and male editors classes respectively), contrary to the male biographies where *men* word is not even among the 100 highest tf-idf scores (the tf-idf scores of word *men* in female editors and male editors class are 0.0234 and 0.0194 respectively).

Looking directly at the tf-idf scores of the words gives a first idea on what are the important words when differentiating texts, but the further aim of investigating about the gender bias of editors pushes for the exploration of differences in importance of words between these groups of editors according to their gender.

Following with the inspection of the *Person* category and after having applied the tests described in section 4.2.3, figure 5.7 shows the confidence intervals for the tf-idf score of all the words considered in this category. Analyzing the women's biographies

in figure 5.7(a), words like *women, woman, married, husband, children* and *father* have a significant higher tf-idf score in female groups of editors versus male groups of editors, indicating that the female contributors tend to give these words more importance than male in women's biographies. Conversely, examining the results for men's articles reveals that, overall, the tf-idf scores for all considered words are lower compared to those in women's articles. Additionally, there are no significant differences in the usage of the considered words between the two groups of editors, except for the words *man* and *married*. Male editors tend to assign more importance to the word *man*, whereas female contributors emphasize the word *married* more.

Next, by examining other categories in more detail, we can observe different results. For example in the category *Athlete* (which is a category that mainly edit the male contributors) showed in figure 5.8, the tf-idf score for the word *women* is significantly greater than the rest of the words considered in this study in women biographies. However, there is no difference in the group of editors being female or male. The word *men* is as well of big importance in the male biographies, but this time there is a difference on how much importance the female editors give to the word compared to the male editors.

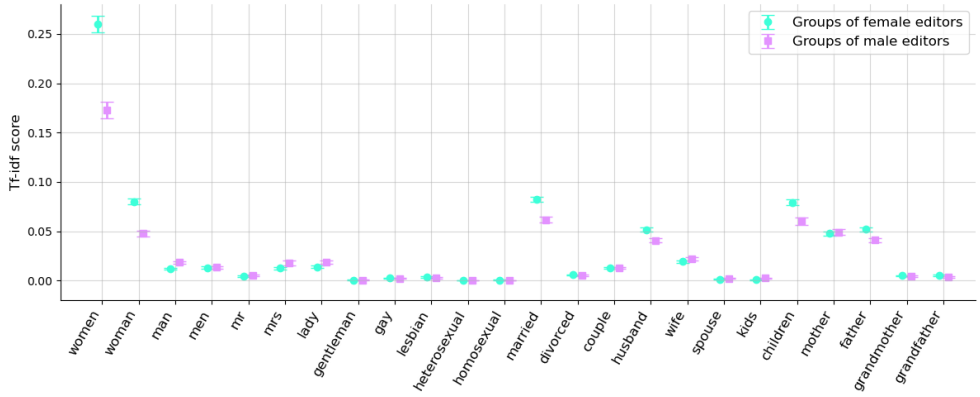
5.2.4 Bigrams of words

As a first approach to studying the bigrams, I analysed the *Person* category. The results of the most 50 frequent bigrams are showed in the tables C.1, C.2, C.3 and C.4.

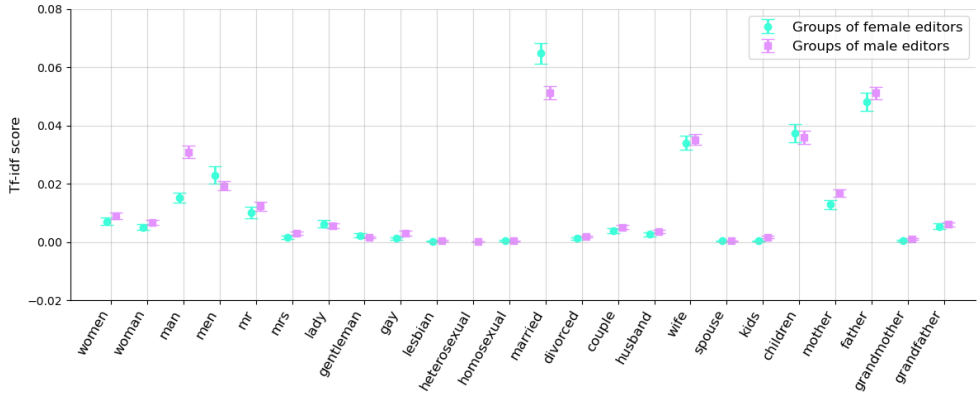
Apart from the usual collocations that one can encounter in English written texts like *New York, personal life, United States, high school, world war, half marathon*, other combinations arose from a first inspection of these results. Gender words like *woman, women* or *female* appear in the most frequent bigrams of the female biographies edited by groups of female contributors (see highlighted bigrams in table C.1). This phenomenon happens only within the female biographies edited by women, and the context in which these gender words are more commonly found are: *first women, first female, women suffrage, woman suffrage, women rights*.

With the aim of comparing the writing styles of women and men, figure 5.9 presents the confidence intervals of the PMI measure described in section 4.2.4 for each of the bigrams considered, differentiating between the pages edited by groups of female versus male contributors.

Reviewing the outcomes showed in the figure above, we can observe that in general the bigrams containing the word *man*, have higher values of PMI compared to the bigrams that include the word *woman*. When it comes to analysing the context in which we find the word *woman*, it is more common to relate this word with the family terms (*family, married, children*) compared to the rest. And it is observable



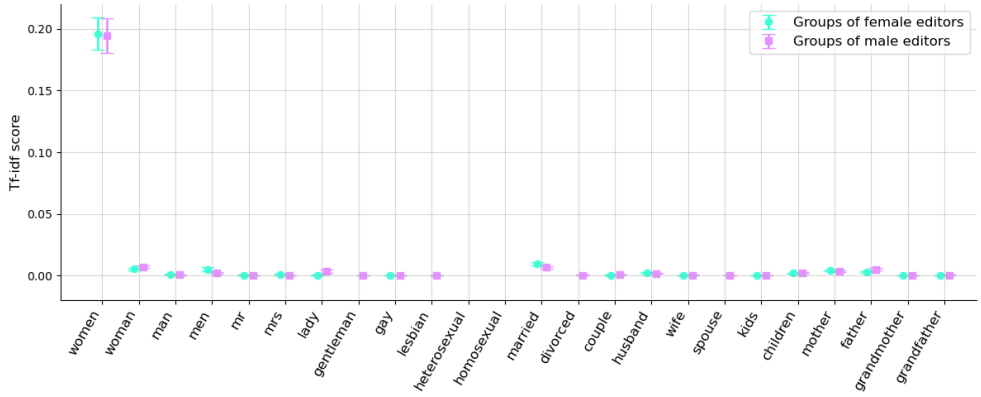
((a)) Female pages in *Person* category.



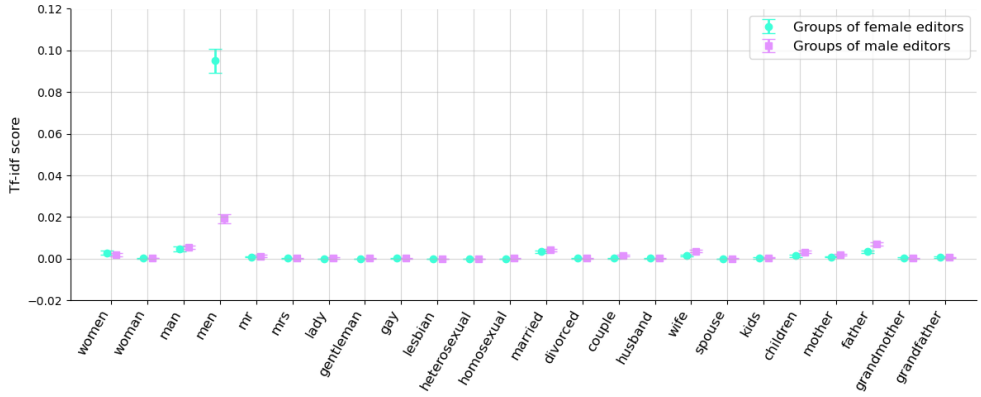
((b)) Male pages in *Person* category.

Figure 5.7: Confidence intervals of tf-idf score for each word, differentiating between the articles edited by female and male groups of editors in the category *Person*. The dot represents the mean of the sampling distribution of the tf-idf score and the error bar the lower and upper bound of the confidence interval. The absence of data for certain words in the figure is because these words are not present in the corresponding articles. Note that the y-axes in both plots have different scales (for a better visualisation).

from the results that the male editors tend to use this combination slightly more often than female. Another key word that combines very frequently with *woman* is the word *rich*. It is noteworthy that this combination has not appeared in the female edited pages, but only men have included this bigram in their texts. In addition, this bigram's PMI score is the one being closer to its male-correspondent (*man, rich*),



((a)) Female pages in *Athlete* category.



((b)) Male pages in *Athlete* category.

Figure 5.8: Confidence intervals of tf-idf score for each word in pages edited by female and male editors groups in the category *Athlete*. The dot represents the mean of the sampling distribution of the tf-idf score and the error bar indicates the lower and upper bound of the confidence interval. The absence of data for certain words in the figure is because these words are not present in the corresponding articles. Note that the y-axes in both plots have different scales (for a better visualisation).

compared to all the other combinations with *woman*.

With respect to the race terms, non of the words analysed (*black, white*) seem to have a very different value of PMI, however, when writing about black men, female contributors tend to accentuate that feature significantly more than contributors.

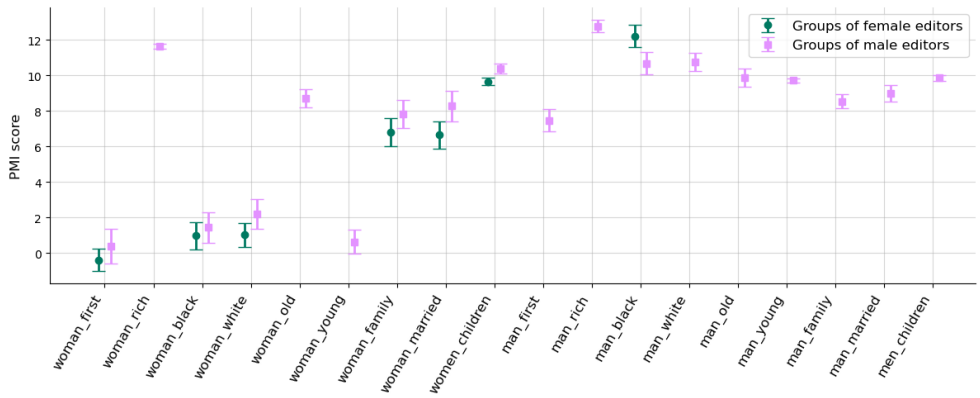


Figure 5.9: Confidence intervals of PMI score for every bigram in *Person* category. Note that the bigrams containing the word *woman* belong to female pages and with the word *man* belong to male pages. The dot represents the mean of the sampling distribution of the PMI score and the error bar indicates the lower and upper bound of the confidence interval. The absence of data for certain bigrams in the figure is because these bigrams are not present in the corresponding set of pages.

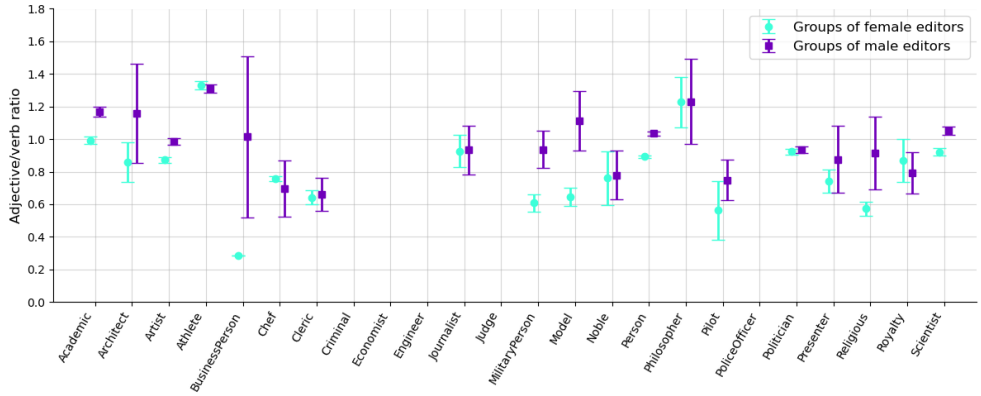
5.2.5 POS-tagging (or Adjective/verb ratio)

As described in 4.2.5, the subjectivity in the language can be measured by the frequency that certain part-of-speech tags appear in the text. In this segment of the research, the results from the adjective-to-verb ratio are displayed. A comparison of all categories divided by the gender of the biographies is presented in figure 5.10.

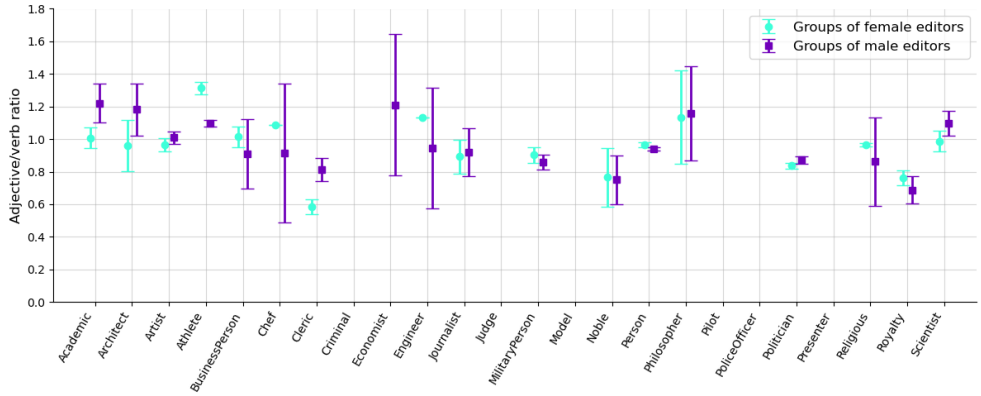
Looking at the *Person* category, which gives us a more general overview when differentiating gender in editors and articles, since it does not contain bias from topic in the articles, it is noticeable that among the female biographies, the adjective to verb ratio is significantly higher in groups of male editors compared to female editors. On the other hand, in the male biographies the difference between these two groups of editors is minimal with a slightly higher ratio from the group of female editors.

Analyzing more generally the female pages, we observe that in none of the categories the adjective to verb ratio is higher within the articles edited by women. A lot of categories do not show significant differences in this measure, but others like *Academic, Artist, MilitaryPerson, Model, Person, Religious, Scientist*, reflect a higher adjective to verb ratio when the text is written by male contributors.

Regarding now the men's biographies, most of the categories have overlapping confidence intervals, indicating that there are not such significant differences in the way female and male contributors write about these persons.



((a)) Female pages



((b)) Male pages

Figure 5.10: Confidence intervals of the adjective/verb ratio, differentiated across groups of female and male editors, for each biographical category. The dot represents the mean of the sampling distribution of the adjective/verb ratio and the error bar indicates the lower and upper bound of the confidence interval. The absence of data for certain categories in the figure is because these categories do not include any pages edited exclusively by groups of female editors or exclusively by groups of male editors.

It is noteworthy that the *Athlete* category is the only one where the ratio of adjectives to verbs is higher among female editors compared to male editors, particularly in articles about male individuals. Furthermore, the values in this category are among the highest for all groups of contributors compared to the rest of categories, especially in articles about female individuals. The fact that the *Athlete* category has among

the highest adjective-to-verb ratios might be due to the fact that describing athletic performance, physical prowess, and personal characteristics often requires the use of adjectives to emphasize the achievements, qualities, and uniqueness of athletes, which are essential aspects of sports narratives.

5.3 Hyperlinks structure

5.3.1 Isolation

In the landscape of articles that have no connecting links to other articles, the results can be seen in figure 5.11. The analysis reveals that in almost every biographical category, the female pages have a higher isolation ratio than the male pages. The isolation ratio range for both genders goes from 0.031 (female pages in *Royalty* category) to 0.988 (female pages in *BusinessPerson* category), indicating that the isolation ratio is heavily dependent on the category. Only in some particular categories like *Model, Royalty, PoliceOfficer* and *Judge*, the male isolated nodes ratio is greater than than in women's biographies. The isolation ratio gender difference can be inspected in detail in figure C.5 in the appendix.

In order to know if these results are significant and not due to chance, we can observe the results of the permutation test applied to the isolation ratio in table C.5 in the appendix. For the majority of categories, the permutation test confirms that we can reject the null hypothesis and thus, conclude that the isolation ratio is significantly different in female articles compared to male. Additionally, it is to note that the isolation ratio is not only "different", but also higher in female pages.

However, there are some specific categories that do not follow this trend: *Athlete, BusinessPerson, Journalist, Judge, Model, Noble, Pilot, PoliceOfficer, Presenter*. The test fail to reject the null hypothesis, indicating no differences is the ratio of isolated nodes between genders.

5.3.2 Homophily

5.3.2.1 Assortativity

Having calculated the assortativity coefficients for all the category networks, the results displayed in figure 5.12 show that there is a variety of assortativity scores dependent on the category. However, most of the categories show a positive assortativity, indicating that nodes tend to connect to other nodes that are similar in the gender attribute.

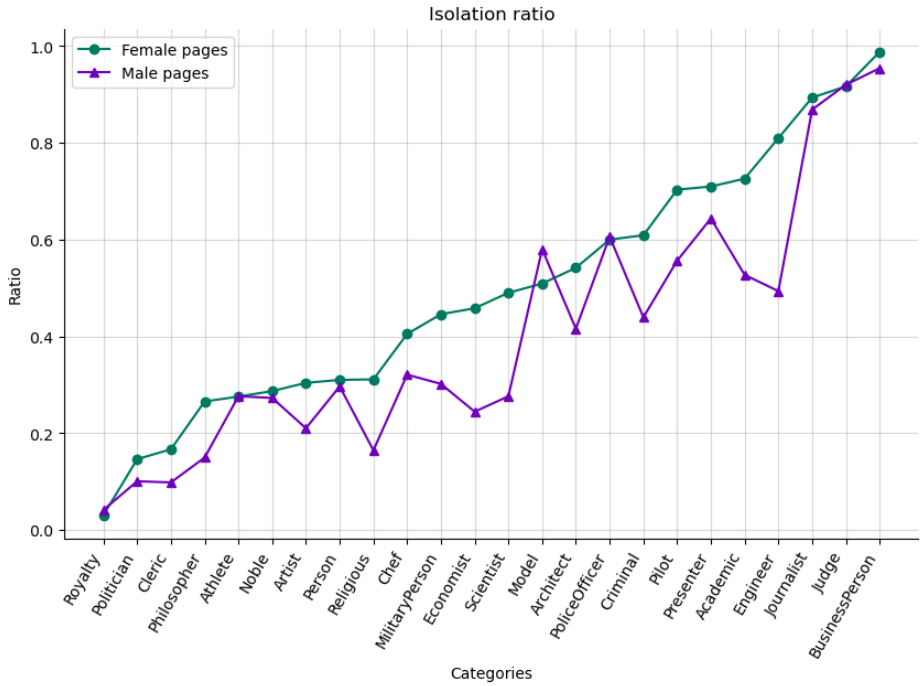


Figure 5.11: Isolation ratio sorted by ascendant values of Female pages isolation ratio.

5.3.2.2 Asymmetry

Regarding the asymmetry of nodes linking to their counter gender, the study discloses that most of the categories showed a positive value of asymmetry A (see formula (4.12)), which suggests that biographies of female individuals are more likely to show a link to articles about male personalities than the other way around.

It is worth noting that the greatest asymmetry is observed in the category *Religious*. Additionally, it is notable that the *Criminal* category is almost the only one exhibiting negative asymmetry, indicating the tendency of men biographies to connect to articles about women.

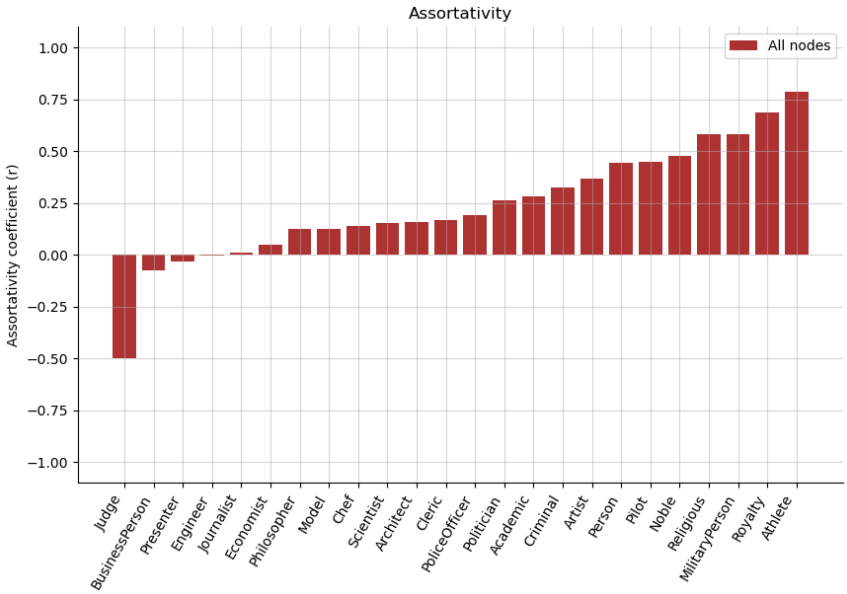


Figure 5.12: Assortativity coefficient per category.

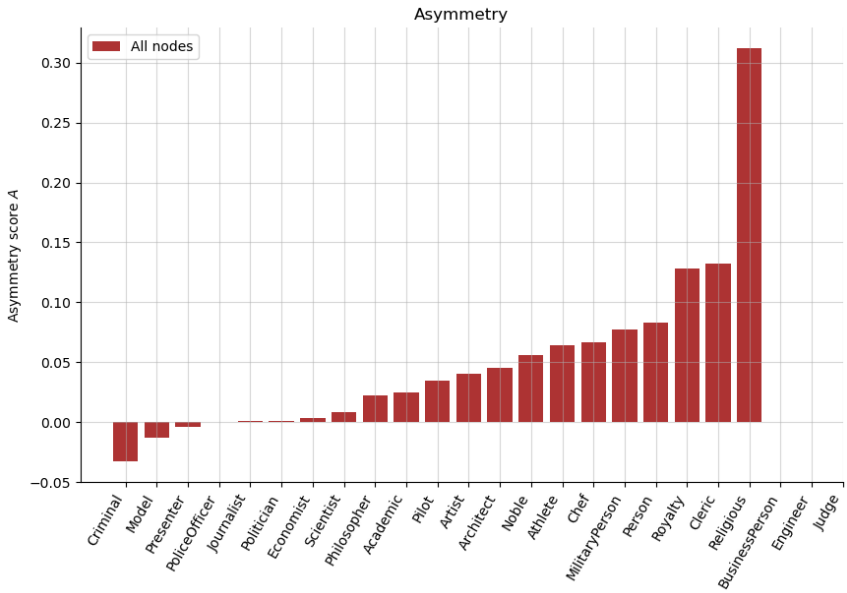


Figure 5.13: Asymmetry score per category. Note that some categories have no values (*BusinessPerson, Engineer* and *Judge*), because there were no edges linking at least one of the four combinations.

CHAPTER 6

Discussion

6.1 Discussion and limitations

6.1.1 Authorship of each article

Due to the fact that Wikipedia is a collaborative online encyclopedia, for some pages, there is a large number of diverse contributors. That is why sometimes it is difficult to set apart the authorship of entire texts or part of the texts.

In this study, the approach taken was the one explained in section 4.1: considering the gender of the group of editors of the article only if all the revisions leading to the last one were made from the same editor gender. While this approach has facilitated the analysis of the gender of the article editors, it has also resulted in a simplification that likely overlooks some nuances and leaves many articles out of the study.

To address the complexity of authorship attribution in collaborative environments like Wikipedia, a detailed revision tracking system can be implemented to identify the extent of each editor's contributions more accurately. Additionally, a method for collaborative authorship attribution can be developed, weighting the contributions based on the number and size of edits, providing a more nuanced understanding of each editor's input.

It is also important to note that this analysis focused solely on a binary gender framework, excluding other forms of non-binary gender expressions. This decision was made due to the insufficient number of articles and contributors identified with these other gender types.

6.1.2 Classification of texts

Performing the initial classification on the articles have been a good initial motivation for the exploration of how gender influences the way that the Wikipedians write and use different lexicon. Also, it has provided some initial hints into what is the important vocabulary used in each specific biography gender, revealing some biases in both content and writing style of the contributors.

6.1.3 Number of words in texts

Comparing the values between women and men biographies, it does not appear to be any significant difference in length between these groups. However, further investigation is warranted to explore this in more detail, as previous studies have found that articles about women tend to be longer than those about men [GLM15; Wag+15].

Still, when examining men's biographies separately, results have showed that female contributors write significantly fewer words than their male counterparts in general. The results obtained in this part could be due to several reasons. One question that may arise is if males tend to create new pages (giving high word counts), and females tend to edit. Considering that the gender of the editors of each page are measured as "totality" (see section 4.1), that means that all the pages considered in this part have been created and edited by the same gender. However, a note has to be made that when obtaining the gender of the editor for each revision, it is possible that not all contributors for a page are considered (because it was not possible to collect their gender).

When it comes to women pages, in general, there were no significant differences in word count between editors' gender. However, it is outstanding the high word count in *Religious* female category. Not only the women editors have higher word count than male editors, but also, it is the higher word count in all groups of pages analysed. Although it is surprising, it is likely due to a single outlier rather than a general rule. In fact, when exploring this group of pages (*Religious* category, women pages and edited by female), we discover that there are only two pages *Mary Reed (missionary)* and *Susie Forrest Swift* both edited by a high amount of contributors (60 and 59 respectively).

This phenomenon gives us the idea that of course the word count is not only influenced by how much the individuals write, but also about how many editors a page has had.

6.1.4 Tf-idf scores

When going to a more lexical level, the results of the tf-idf analysis in women's biographies show that in general female contributors give the words *women, woman, married, husband, children, father* more relevance compared to the male contributors. This could infer that women tend to reinforce the female gender more with these gender-associated words and also focus more on the individual's family aspects, through the family terms.

However, in men's biographies, there is a significant difference in the relevance of the word *man* between the gender of contributors: male contributors tend to consider it more relevant than female contributors. This observation is symmetrical to

the female case.

Furthermore, a direct comparison of the tf-idf scores for the term *woman* in female articles and *man* in male articles reveals that *woman* holds significantly greater relevance. This is not an unexpected result, considering how the male gender has been frequently seen as the "default" gender [CM20], meaning that, historically or traditionally, male perspectives or experiences have been considered the default or standard in research, often due to historical biases and underrepresentation of other genders.

6.1.5 Bigrams of words

In the study of bigrams, it was observed that bigrams containing the word *man* have higher PMI values compared to those including the word *woman*. This finding may seem counterintuitive given that female-gender-related words are generally more prominent when analyzed individually (results of part 5.2.3). A potential reason for this is that PMI has a well-known tendency to give higher scores to low-frequency events [RN11], instead of what would be more preferable for this application which is to give a higher score for pairs of words whose relatedness is supported by more evidence.

The observation that male editors more frequently associate the word *woman* with family terms like *family, married, children* compared to female editors could stem from entrenched societal norms and perceptions. Male editors may adhere to traditional gender roles that emphasize women's roles within the family, reflecting cultural and social influences. These editors might perceive women primarily through the lens of familial responsibilities, while female editors may focus on a wider spectrum of women's lives, such as career achievements, personal accomplishments, or individual identity.

The frequent combination of *rich* with *woman*, which appears only in pages edited by men, might be due to gendered perspectives on wealth and cultural biases that lead men to highlight financial status more prominently. Men may perceive wealth as a significant characteristic of women, reflecting societal stereotypes, while female contributors might focus on other aspects like career or achievements. Men might have traditionally been more associated with financial matters and thus may bring that perspective into their writing more frequently.

Female contributors emphasize the racial identity of black men more than their male counterparts, despite the overall association scores for race-related terms not differing significantly.

These findings confirm that men and women are indeed presented differently on Wikipedia as previous studies have also demonstrated [Wag+15; Wag+16]. These differences may stem from historical gender inequalities, such as the greater challenges women faced in achieving fame due to unequal access to resources and the fact that history has predominantly been documented from a male perspective.

6.1.6 POS-tagging (or adjective/verb ratio)

In the case of women's biographies, the higher adjective-to-verb ratio in articles written by men, compared to those by female editors, might suggest that men use more abstract language when describing women. According to the study by Anne Maass et al. [Maa+89], this could imply that men, when describing the out-group in a more abstract manner, might be emphasizing undesirable attributes of women. In contrast, for men's biographies, the adjective-to-verb ratio used by female and male contributors does not show significant differences.

There could be some potential explanations for these results like the gender dynamics influencing how women feel about describing men and vice versa. Female editors may strive for neutrality to avoid potential backlash or accusations of bias, resulting in a language use similar to their male counterparts. There might be less stereotyping and fewer preconceived notions about men compared to women, leading both genders to use similar descriptive language.

6.1.7 Isolation ratios in hyperlinks network

The hyperlink structure in Wikipedia is a crucial framework for article retrieval [KK08]. Examining the isolation ratios of female and male Wikipedia pages reveals that women's articles are consistently more isolated than men's. However, there are some page categories where the difference is minimal or non-existent.

The case of the category *Athlete* is quite surprising. Despite this category not being particularly representative of women's biographies and having relatively few contributions from female editors, the permutation test fail to reject the hypothesis that the isolation ratios of the two gender editor groups are the same. One possible explanation for this phenomenon is that this category is probably very symmetrical in terms of gender, since in most sport disciplines there are male and female competing. This could give the network of *Athlete* pages a more gender-symmetrical structure in the way pages link to each other.

6.1.8 Homophily of hyperlinks network

The results on the homophily of the categories articles' networks have showed that it is clear that there is a gender gap when it comes to where the articles connect to. The

positive values of the assortativity coefficients indicate that there is not a considerable connectivity between the two gender pages groups considered. The asymmetry coefficient has given us an idea on where the biographies in Wikipedia point to: tendency of articles linking to male biographies.

It is important to note the inverse tendency of asymmetry observed in the *Criminal* category, meaning that articles about men tend to link more to articles about women than the opposite. Additionally, the tone associated with this category is frequently and predominantly negative, reflecting its association with crime, legal issues, and other adverse events or behaviors. So, maybe this is not a category to be proud of being linked to?

The above results show the existence of assortativity and asymmetry across genders controlling for degree. However, structural biases can also manifest in the centrality measures, as suggested by *the Smurfette principle* [Pol91].

6.2 Future research

In light of the results and limitations discussed in previous sections, several opportunities for future research arise:

A direct opportunity for future research is to investigate the different language editions of Wikipedia pages. Contributors writing in different languages might exhibit other language bias and trends. Also, different language editions might cover unique topics or details not present in others, so the page categories distribution might vary. This approach could also reveal how different cultures present and prioritize information, providing a richer understanding of cultural perspectives and biases.

Analyzing the difference between consecutive revisions of the same article could also give great insight in how each user contributes into the editing process. What would happen if a single woman makes an edit into a page that was previously edited only by men? Also, this could help for the study of the interaction between different contributors, revealing collaborative or conflicting editing behaviors.

Further investigations could differentiate between the birth years of biography subjects. There might be differences in patterns of participation and attention given by Wikipedia editors to more recent subjects, due to the increased volume of information available about them. This suggestion is inspired by Wagner et al. (2015) [Wag+15], which studied inequalities in Wikipedia articles by differentiating biographies of subjects born pre-1900 and 1900-present.

This study was solely based on a single human characteristic: gender. However, it is of great importance and interest to consider other features such as age, nationality,

socio-economic status, etc. It could be beneficial because it would provide a more comprehensive understanding of the factors influencing the the content in Wikipedia biographies. Considering multiple characteristics allows for a richer analysis, revealing complex interactions and correlations that may not be evident when focusing only on the gender variable.

Lastly, it remains open whether these trends and biases are unique to Wikipedia contributors. Future research should explore whether the biases in Wikipedia text content reflect those present in general media, or if they are influenced further by the specific demographics and perspectives of the Wikipedia editor community.

CHAPTER 7

Conclusion

Wikipedia biographies still display a significant gender imbalance among contributors and pages. Despite various efforts to close the gender gap, such as women-only edit-a-thons, *Women in Red*, and other organizations, the male gender continue to dominate in both contributions and biographies. Achieving equality in gender representation is a long journey, as historically, male achievements and perspectives have dominated over women's.

This research conclusively unveils a gender gap across several dimensions in Wikipedia: first, the unbalanced distribution of gender among contributors and biographies; second, the disparities in language usage when editing texts between male and female contributors and articles; and finally, the differences in the connectivity of articles based on gender.

The initial classification of articles through ML models highlighted gender influences in writing styles and lexicon, revealing some biases. This provided an initial good motivation for keeping exploring in depth these biases.

When immersing into the text analysis, it was revealed that female contributors generally write fewer words than male counterparts in men's biographies. Analyzing the lexical aspect, the conclusion is that in women's biographies, women tend to reinforce the female gender with the use of gender-associated words and also they focus more on the individual's family aspects for both men and women's biographies. As for the context in which the genders are portrayed, it was revealed that male editors often associate *woman* with family terms, reflecting societal norms, while female editors focus on a broader range of women's lives. The more abstract language employed by men when describing female personalities could imply that men emphasize undesirable attributes of women.

The connectivity of the hyperlinks network is an important feature in information retrieval for Wikipedia and through this analysis it has been revealed that women's articles are generally more isolated than men's, implying a block in the exploration of women personalities through this network. Additionally, a clear gender gap in article connectivity is evident. First, there is limited connectivity between male and female biography pages. Second, there is a tendency for women's articles to link to male

biographies, but not the reverse.

In conclusion, these findings underscore the persistent and multifaceted gender gaps on one of the world's most influential digital platforms. These findings highlight the need for balanced gender representation in Wikipedia editing, both in biographies and among contributors. Achieving this balance could lead to a less biased portrayal of personalities through a more unified use of language and importance of certain topics.

Bibliography

[And17] Jannis Androutsopoulos. "Online data collection." In: *Data collection in sociolinguistics*. Routledge, 2017, pages 233–244.

[Ant+11] Judd Antin et al. "Gender differences in Wikipedia editing." In: *Proceedings of the 7th international symposium on wikis and open collaboration*. 2011, pages 11–14.

[Arg+03] Shlomo Argamon et al. "Gender, genre, and writing style in formal written texts." In: *Text & talk* 23.3 (2003), pages 321–346.

[Arg+07] Shlomo Argamon et al. "Mining the blogosphere: Age, gender and the varieties of self-expression." In: *First Monday* (2007).

[Arg+09] Shlomo Argamon et al. "Automatically profiling the author of an anonymous text." In: *Communications of the ACM* 52.2 (2009), pages 119–123.

[AS18] Rajesh Arumugam and Rajalingappaa Shanmugamani. *Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications*. Packt Publishing Ltd, 2018.

[AS23] Ali Acilar and Øystein Sæbø. "Towards understanding the gender digital divide: A systematic literature review." In: *Global knowledge, memory and communication* 72.3 (2023), pages 233–249.

[AV01] James Andreoni and Lise Vesterlund. "Which is the fair sex? Gender differences in altruism." In: *The Quarterly Journal of Economics* 116.1 (2001), pages 293–312.

[BC16] Julia B Bear and Benjamin Collier. "Where are the women in Wikipedia? Understanding the different psychological experiences of men and women in Wikipedia." In: *Sex roles* 74 (2016), pages 254–265.

[Bea+09] Lori Beaman et al. "Powerful women: does exposure reduce bias?" In: *The Quarterly journal of economics* 124.4 (2009), pages 1497–1540.

[BES14] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. "Gender identity and lexical variation in social media." In: *Journal of Sociolinguistics* 18.2 (2014), pages 135–160.

[Bey+22] Pablo Beytía et al. “Visual gender biases in wikipedia: A systematic evaluation across the ten most spoken languages.” In: *Proceedings of the International AAAI Conference on Web and Social Media*. Volume 16. 2022, pages 43–54.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009.

[BLS21] Sebastian Brückner, Florian Lemmerich, and Markus Strohmaier. “Inferring sociodemographic attributes of Wikipedia editors: State-of-the-art and implications for editor privacy.” In: *Companion Proceedings of the Web Conference 2021*. 2021, pages 616–622.

[BO05] Constantinos Boulis and Mari Ostendorf. “A quantitative analysis of lexical differences between genders in telephone conversations.” In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*. 2005, pages 435–442.

[Bor06] Lera Boroditsky. “Linguistic relativity.” In: *Encyclopedia of cognitive science* (2006).

[Bor11] Lera Boroditsky. “How language shapes thought.” In: *Scientific American* 304.2 (2011), pages 62–65.

[BP10] Hoda Baytiyeh and Jay Pfaffman. “Volunteers in Wikipedia: Why the community matters.” In: *Journal of Educational Technology & Society* 13.2 (2010), pages 128–140.

[BP22] Jonah Berger and Grant Packard. “Using natural language processing to understand people and culture.” In: *American Psychologist* 77.4 (2022), page 525.

[BS12] Dennis D Boos and Leonard A Stefanski. *Essential statistical inference: theory and methods*. Volume 120. Springer, 2012, pages 413–448.

[BW22] Pablo Beytía and Claudia Wagner. “Visibility layers: A framework for facing the complexity of the gender gap in wikipedia content.” In: *Internet Policy Review* (2022).

[Cab+18] Benjamin Cabrera et al. “The gender gap in Wikipedia talk pages.” In: *Proceedings of the international AAAI conference on web and social media*. Volume 12. 1. 2018.

[CB12] Benjamin Collier and Julia Bear. “Conflict, confidence, or criticism: An empirical examination of the gender gap in Wikipedia.” In: *Proceedings of the ACM 2012. Conference on Computer Supported Cooperative Work, New York*. 2012, pages 383–392.

[CCC10] Hichang Cho, MeiHui Chen, and Siyoung Chung. “Testing an integrative theoretical model of knowledge-sharing behavior in the context of Wikipedia.” In: *Journal of the American Society for Information Science and Technology* 61.6 (2010), pages 1198–1212.

[CFD17] Zhihui Cai, Xitao Fan, and Jianxia Du. "Gender and attitudes toward technology use: A meta-analysis." In: *Computers & Education* 105 (2017), pages 1–13.

[CH90] Kenneth Ward Church and Patrick Hanks. "Word Association Norms, Mutual Information, and Lexicography." In: *Computational Linguistics* 16.1 (1990), pages 22–29. URL: https://aclanthology.org/J90-1003.

[Cha21] Matt Chase. "Wikipedia is 20, and its reputation has never been higher." In: *The Economist* (January 2021). Accessed: 2024-05-23. URL: https://www.economist.com/international/2021/01/09/wikipedia-is-20-and-its-reputation-has-never-been-higher.

[CM20] Sapna Cheryan and Hazel Rose Markus. "Masculine defaults: Identifying and mitigating hidden cultural biases." In: *Psychological Review* 127.6 (2020), page 1022.

[Coh11] Noam Cohen. "Define gender gap? Look up Wikipedia's contributor list." In: *The New York Times* 30.01 (2011).

[Con18] Intersoft Consulting. *GDPR*. Accessed: 2024-07-01. 2018. URL: https://gdpr-info.eu/.

[CP98] Jennifer Coates and Pia Pichler. "Language and gender." In: *A Reader (2nd ed.) Oxford/Malden: Wiley* (1998).

[Dix+14] Laura J Dixon et al. "Gendered space: The digital divide between male and female users in internet public access sites." In: *Journal of Computer-Mediated Communication* 19.4 (2014), pages 991–1009.

[ED18] Organisation for Economic Co-operation and Development. *BRIDGING THE DIGITAL GENDER DIVIDE: INCLUDE, UPSKILL, INNOVATE*. Accessed: 2024-06-17. 2018. URL: https://www.oecd.org/digital/bridging-the-digital-gender-divide.pdf.

[ES13] Stine Eckert and Linda Steiner. *Wikipedia's Gender Gap. Media (dis) parity: A gender battleground*. 2013.

[ET94] Bradley Efron and Robert Tibshirani. "Bootstrapping Regression Models." In: *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall/CRC, 1994. Chapter 21, pages 327–349.

[FC19] Mark M Fredrickson and Yuguo Chen. "Permutation and randomization tests for network analysis." In: *Social Networks* 59 (2019), pages 171–183.

[Fer+21] Núria Ferran-Ferrer et al. "The gender gap on the Spanish Wikipedia: Listening to the voices of women editors." In: *Profesional de la información* 30.5 (2021).

[FH61] Robert M Fano and David Hawkins. "Transmission of information: A statistical theory of communications." In: *American Journal of Physics* 29.11 (1961), pages 793–794.

[Fin13] A Finkbeiner. *What I'm not going to do: Do media have to talk about family matters*. 2013.

[FJR06] Mary Frank Fox, Deborah G Johnson, and Sue V Rosser. *Women, gender, and technology*. University of Illinois Press, 2006.

[FKW18] Masoomali Fatehkia, Ridhi Kashyap, and Ingmar Weber. "Using Facebook ad data to track the global digital gender gap." In: *World Development* 107 (2018), pages 189–209.

[Fou20] World Wide Web Foundation. *Women's Rights Online: closing the digital gender gap for a more equal world*. Accessed: 2024-05-27. 2020. URL: https://webfoundation.org/research/womens-rights-online-2020/.

[Fou24] Wikimedia Foundation. *Wikimedia Foundation Mission*. Accessed: 2024-05-30. 2024. URL: https://wikimediafoundation.org/about/mission/.

[Fra23] John M Francis. "The Relationship between Language and Thought: How Does Language Shape Human Perception of the World?" In: *Literature and Linguistics Journal* 2.2 (2023), pages 12–19.

[FW17] Heather Ford and Judy Wajcman. "'Anyone can edit', not everyone does: Wikipedia's infrastructure and the gender gap." In: *Social studies of science* 47.4 (2017), pages 511–527.

[FWD19] Gemma Fitzsimmons, Mark J Weal, and Denis Drieghe. "The impact of hyperlinks on reading text." In: *PLoS One* 14.2 (2019), e0210900.

[GD15] Ananya Goswami and Sraboni Dutta. "Gender differences in technology usage—A literature review." In: *Open Journal of Business and Management* 4.1 (2015), pages 51–59.

[GLM15] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. "First women, second sex: Gender bias in Wikipedia." In: *Proceedings of the 26th ACM conference on hypertext & social media*. 2015, pages 165–174.

[GS17] Maude Gauthier and Kim Sawchuk. "Not notable enough: feminism and expertise in Wikipedia." In: *Communication and critical/cultural studies* 14.4 (2017), pages 385–402.

[GSG10] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. "Wikipedia survey–overview of results." In: *United Nations University: Collaborative Creativity Group* 8 (2010), pages 1158–1178.

[GSR09] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. "Stylometric analysis of bloggers' age and gender." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Volume 3. 1. 2009, pages 214–217.

[Gua14] The Guardian. *The Guardian view on Wikipedia: evolving truth*. Accessed: 2024-07-15. 2014. URL: https://www.theguardian.com/commentisfree/2014/aug/07/guardian-view-wikipedia-evolving-truth.

[Her08] SC Herring. *Gender and power in on-line communication. The handbook of language and gender/ed. by J. Holmes and M. Meyerhoff*. 2008.

[Hin19] Marit Hinnosaar. "Gender inequality in new media: Evidence from Wikipedia." In: *Journal of economic behavior & organization* 163 (2019), pages 262–276.

[HS13] Benjamin Mako Hill and Aaron Shaw. "The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation." In: *PloS one* 8.6 (2013), e65782.

[HS15] Eszter Hargittai and Aaron Shaw. "Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia." In: *Information, communication & society* 18.4 (2015), pages 424–442.

[Hum24] Humaniki. *Humaniki. Search gender metrics*. Accessed: 2024-05-31. 2024. URL: `https://humaniki.wmcloud.org/search`.

[Hyn18] Mike Hynes. "Shining a brighter light into the digital 'black box': A call for stronger sociological (re) engagement with digital technology design, development and adoption debates." In: *Irish Journal of Sociology* 26.1 (2018), pages 94–126.

[Jac+01] Linda A Jackson et al. "Gender and the Internet: Women communicating and men searching." In: *Sex roles* 44 (2001), pages 363–379.

[Jem20] Dariusz Jemielniak. *Common knowledge? An ethnography of Wikipedia*. Stanford University Press, 2020.

[Kas+20] Ridhi Kashyap et al. "Monitoring global digital gender inequality using the online populations of Facebook and Google." In: *Demographic Research* 43 (2020), pages 779–816.

[KK08] Jaap Kamps and Marijn Koolen. "The importance of link evidence in Wikipedia." In: *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30*. Springer. 2008, pages 270–282.

[KK15] Max Klein and Piotr Konieczny. "Wikipedia in the world of global gender inequality indices: What the biography gender gap is measuring." In: *Proceedings of the 11th International Symposium on Open Collaboration*. 2015, pages 1–2.

[KK21] Anna Karczewska and Katarzyna Kukowska. "Cultural dimension of femininity: masculinity in virtual organizing knowledge sharing." In: *European Conference on Knowledge Management*. Academic Conferences International Limited. 2021, pages 414–422.

[Koc20] Richie Koch. *What is considered personal data under the EU GDPR?* Accessed: 2024-07-01. 2020. URL: `https://gdpr.eu/eu-gdpr-personal-data/`.

[KR96] Judith L Klavans and Philip Resnik. *The balancing act: combining symbolic and statistical approaches to language*. Volume 32. MIT press, 1996.

[Lam+11] Shyong (Tony) K Lam et al. "WP: clubhouse? An exploration of Wikipedia's gender imbalance." In: *Proceedings of the 7th international symposium on Wikis and open collaboration*. 2011, pages 1–10.

[Lan+12] David Laniado et al. "Emotions and dialogue in a peer-production community: the case of Wikipedia." In: *proceedings of the eighth annual international symposium on wikis and open collaboration*. 2012, pages 1–10.

[LB19] Victoria Leonard and Sarah E Bond. "Advancing feminism online: Online tools, visibility, and women in classics." In: *Studies in late antiquity* 3.1 (2019), pages 4–16.

[LR09] Shyong (Tony) K Lam and John Riedl. "Is Wikipedia growing a longer tail?" In: *Proceedings of the 2009 ACM International Conference on Supporting Group Work*. 2009, pages 105–114.

[Luy12] Brendan Luyt. "The inclusivity of Wikipedia and the drawing of expert boundaries: An examination of talk pages and reference lists." In: *Journal of the American Society for Information Science and Technology* 63.9 (2012), pages 1868–1878.

[Maa+89] Anne Maass et al. "Language use in intergroup contexts: The linguistic intergroup bias." In: *Journal of personality and social psychology* 57.6 (1989), page 981.

[Med24] MediaWiki. *MediaWiki API Main page*. Accessed: 2024-02-22. 2024. URL: `https://www.mediawiki.org/wiki/API:Main_page`.

[MEP19] Amanda Menking, Ingrid Erickson, and Wanda Pratt. "People who can take it: How women Wikipedians negotiate and navigate safety." In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pages 1–14.

[Min+21] Julià Minguillón et al. "Exploring the gender gap in the Spanish Wikipedia: Differences in engagement and editing practices." In: *PLoS one* 16.2 (2021), e0246702.

[Mor+13] Jonathan T Morgan et al. "Tea and sympathy: crafting positive new user experiences on wikipedia." In: *Proceedings of the 2013 conference on Computer supported cooperative work*. 2013, pages 839–848.

[New03] Mark EJ Newman. "Mixing patterns in networks." In: *Physical review E* 67.2 (2003), page 026126.

[Ngu+13] Dong Nguyen et al. "" How old do you think I am?" A study of language and age in Twitter." In: *Proceedings of the international AAAI conference on web and social media*. Volume 7. 1. 2013, pages 439–448.

[Ngu+16] Dong Nguyen et al. “Computational sociolinguistics: A survey.” In: *Computational linguistics* 42.3 (2016), pages 537–593.

[NO06] Scott Nowson and Jon Oberlander. “The Identity of Bloggers: Openness and Gender in Personal Weblogs.” In: *AAAI spring symposium: Computational approaches to analyzing weblogs*. Palo Alto, CA. 2006, pages 163–167.

[OGR08] Felipe Ortega, Jesus M Gonzalez-Barahona, and Gregorio Robles. “On the inequality of contributions to Wikipedia.” In: *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*. IEEE. 2008, pages 304–304.

[Oko+12] Chitu Okoli et al. “The people’s encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia.” In: *Available at SSRN 2021326* (2012).

[OS18] Amy O’Donnell and Caroline Sweetman. *Introduction: gender, development and ICTs*. 2018.

[Ott10] Jahna Otterbacher. “Inferring gender of movie reviewers: exploiting writing style, content and metadata.” In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pages 369–378.

[Ott15] Jahna Otterbacher. “Linguistic bias in collaboratively produced biographies: crowdsourcing social stereotypes?” In: *Proceedings of the International AAAI Conference on Web and Social Media*. Volume 9. 1. 2015, pages 298–307.

[Pan11] Mani Pande. “Shedding light on women who edit Wikipedia.” In: *Wikimedia Blog* (2011).

[Pic+21] Tiziano Piccardi et al. “On the Value of Wikipedia as a Gateway to the Web.” In: *Proceedings of the Web Conference 2021*. 2021, pages 249–260.

[Pol91] Katha Pollitt. “Hers; the Smurfette principle.” In: *The New York Times* 7.04 (1991).

[Qai+22] Farah Qaiser et al. “How academic institutions can help to close Wikipedia’s gender gap.” In: *Nature* (2022).

[Red+20] Miriam Redi et al. “A taxonomy of knowledge gaps for wikimedia projects (second draft).” In: *arXiv preprint arXiv:2008.12314* (2020).

[RN11] François Role and Mohamed Nadif. “Handling the impact of low frequency events on co-occurrence based measures of word similarity-a case study of pointwise mutual information.” In: *International Conference on Knowledge Discovery and Information Retrieval*. Volume 2. Scitepress. 2011, pages 218–223.

[Rob22] Brandon Andrew Robinson. "Non-binary embodiment, queer knowledge production, and disrupting the Cisnormative field: Notes from a trans ethnographer." In: *The Journal of Men's Studies* 30.3 (2022), pages 425–445.

[Sch+06] Jonathan Schler et al. "Effects of age and gender on blogging." In: *AAAI spring symposium: Computational approaches to analyzing weblogs*. Volume 6. 2006, pages 199–205.

[Sch+13] H Andrew Schwartz et al. "Personality, gender, and age in the language of social media: The open-vocabulary approach." In: *PloS one* 8.9 (2013), e73791.

[Sem24] Semrush. *Semrush top websites*. Accessed: 2024-05-23. 2024. URL: `https://www.semrush.com/website/top/`.

[SH09] Joachim Schroer and Guido Hertel. "Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it." In: *Media Psychology* 12.1 (2009), pages 96–120.

[Sin01] Sameer Singh. "A pilot study on gender differences in conversational speech on lexical richness measures." In: *Literary and Linguistic Computing* 16.3 (2001), pages 251–264.

[SK20] Sabrina Sobieraj and Nicole C Krämer. "Similarities and differences between genders in the usage of computer with different levels of technological complexity." In: *Computers in Human Behavior* 104 (2020), page 106145.

[Spa72] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval." In: *Journal of documentation* 28.1 (1972), pages 11–21.

[Sta24] Internet World Stats. *Internet World Stats*. Accessed: 2024-05-24. 2024. URL: `https://www.internetworldstats.com/`.

[Sta78] JE Staddon. "Theory of behavioral power functions." In: *Psychological Review* 85.4 (1978), page 305.

[Ste17] R Stephenson-Goodknight. *Gender diversity mapping project–Diversity Conference 2017*. 2017.

[SY14] Anna Samoilenko and Taha Yasseri. "The distorted mirror of Wikipedia: a quantitative analysis of Wikipedia coverage of academics." In: *EPJ data science* 3 (2014), pages 1–11.

[Tho92] Gillian Thomas. *A position to command respect: women and the eleventh Britannica*. Scarecrow Press, 1992.

[Tor23] Joana Soler Torramilans. "The Gender Gap in English Wikipedia: An Examination of Participation, Content Preferences, and Collaborative Practices." Master's thesis. University of Copenhagen, August 2023.

[Tri23] Francesca Tripodi. “Ms. Categorized: Gender, notability, and inequality on Wikipedia.” In: *New media & society* 25.7 (2023), pages 1687–1707.

[Ukw+21] Scholastica Chizoma Ukwoma et al. “Unveiling the veiled: Wikipedia collaborating with academic libraries in Africa in creating visibility for African women through Art+ Feminism Wikipedia edit-a-thon.” In: *Digital library perspectives* 37.4 (2021), pages 449–462.

[Vit17] Marie A Vitulli. *Writing women in mathematics into Wikipedia*. 2017.

[Wag+15] Claudia Wagner et al. “It’s a man’s Wikipedia? Assessing gender inequality in an online encyclopedia.” In: *Proceedings of the international AAAI conference on web and social media*. Volume 9. 1. 2015, pages 454–463.

[Wag+16] Claudia Wagner et al. “Women through the glass ceiling: gender asymmetries in Wikipedia.” In: *EPJ data science* 5 (2016), pages 1–24.

[Wik11] Wikimedia. *Wikipedia editors study: results from the editor survey, April 2011*. Accessed: 2024-06-20. 2011. URL: `https://upload.wikimedia.org/wikipedia/commons/7/76/Editor_Survey_Report_-_April_2011.pdf`.

[Wik24a] Wikimedia. *Community insights/2018 report*. Accessed: 2024-06-20. 2024. URL: `https://meta.wikimedia.org/wiki/Community_Insights/2018_Report`.

[Wik24b] Wikipedia. *Wikipedia: Wikipedians*. Accessed: 2024-05-30. 2024. URL: `https://en.wikipedia.org/wiki/Wikipedia:Wikipedians`.

[Wik24c] Wikipedia. *WikiProject Women in Red*. Accessed: 2024-07-01. 2024. URL: `https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Women_in_Red`.

[WZ18] Jess Wade and Maryam Zaringhalam. “Why we’re editing women scientists onto Wikipedia.” In: *Nature* 14 (2018).

[YWK16] Amber Young, Ari D Wigdor, and Gerald Kane. “It’s not what you think: Gender bias in information about Fortune 1000 CEOs on Wikipedia.” In: (2016).

[ZFW17] Olga Zagovora, Fabian Flöck, and Claudia Wagner. “”(Weitergeleitet von Journalistin)” The Gendered Presentation of Professions on Wikipedia.” In: *Proceedings of the 2017 ACM on web science conference*. 2017, pages 83–92.

APPENDIX A

Data processing

A.1 Wikitext example of a page and revision

```
{{Use dmy dates|date=March 2020}}
{{use British English|date=August 2016}}
{{BLP sources|date=March 2011}}
{{Infobox person
|name              = Raymond Snoddy<ref name=BBC_Snoddy>{{cite web|date=1
    ↪ October 2004|title=NewsWatch &#124; Profiles &#124; Raymond Snoddy|
    ↪ url=http://news.BBC.co.uk/newswatch/ifs/hi/newsid_3700000/
    ↪ newsid_3701800/3701840.stm|website=news.BBC.co.uk|publisher=[[BBC
    ↪ News]] / [[BBC Online]]|accessdate=28 August 2016}}</ref>
|honorific_suffix = [[Order of the British Empire|OBE]]<ref name=
    ↪ LondonGazette/>
|image             =  <!--filename only, no 'File:' or 'Image:' prefix, and
    ↪ no enclosing [[brackets]]-->
|image_size        = <!--DISCOURAGED per WP:IMGSIZE; use image_upright-->
|alt               =
|caption           =
|birth_name        = Matthew Raymond Snoddy<ref name=LondonGazette/>
|birth_date        = {{birth year and age|1946}}
|birth_place       = [[Larne]], [[County Antrim]], Northern Ireland
|nationality       = [[British people|British]]
|citizenship       =
|education         =
|alma_mater        = [[Queen's University, Belfast]]<ref name=BBC_Snoddy/>
|occupation        = [[Broadcast journalism|Broadcast]] and print media [[
    ↪ News media|news]] [[journalist]]
|years_active      = [[Wiktionary:circa|c.]]1970 - present
|employer          = [[BBC News]] <small>(former)</small>
|organization      =
|known_for         = [[News broadcasting#Television news|Television news]] [[
    ↪ News broadcasting|presenter]], [[author]]
|notable_works     = [[Michael Green (television magnate)|Michael Green]]
    ↪ biography: ''The Good, the Bad and the Unacceptable: The Hard News
    ↪ about the British Press''<ref name=BBC_Snoddy/> <!--produces label '
    ↪ Notable work'; may be overridden by |credits=, which produces label '
    ↪ Notable credit(s)'; or by |works=, which produces label 'Works'-->
|style             =
|height            =  <!--'X cm', 'X m' or 'X ft Y in', plus optional
    ↪ reference (conversions are automatic)-->
|television        =
```

```
|title            = Presenter of [[BBC]] [[BBC News (TV channel)|News 24]] '
    ↪ '[[NewsWatch (TV series)|NewsWatch]]''<ref name=BBC_Snoddy/>
|term             = 2004 - 2013
|predecessor      = None
|successor        = [[Samira Ahmed]]
|boards           =
|spouse           = Diana<ref name=Guardian/>
|partner          =  <!--unmarried long-term partner-->
|children         =
|parents          =  <!--overrides mother and father parameters-->
|mother           =  <!--may be used (optionally with father parameter) in
    ↪ place of parents parameter (displays 'Parent(s)' as label)-->
|father           =  <!--may be used (optionally with mother parameter) in
    ↪ place of parents parameter (displays 'Parent(s)' as label)-->
|relatives        =
|family           =
|awards           =
|website          =  <!--{{URL|example.com}}-->
|footnotes        =
}}
'''Matthew Raymond Snoddy [[Order of the British Empire|OBE]]''',<ref name=
    ↪ LondonGazette>{{London Gazette|supp=y|issue=55710|page=16|date=31
    ↪ December 1999}}</ref> born {{birth year and age|1946}}, commonly
    ↪ known as '''Raymond Snoddy''', is a British [[news media]] [[
    ↪ journalist]], [[News broadcasting#Television news|television]] [[News
    ↪  broadcasting|presenter]], author and [[Pundit|media commentator]].
    ↪ From its inception in 2004, until January 2013, he was the original
    ↪ and sole presenter of the [[BBC]] [[BBC News (TV channel)|News 24's]]
    ↪  weekly viewer [[Right of reply#By editorial policy|right-to-reply]]
    ↪ programme ''[[NewsWatch (TV series)|NewsWatch]]''.<ref name=
    ↪ BBC_Snoddy/><ref name=Guardian>{{cite news|editor=Jason Deans|date=10
    ↪  August 2004|title=Snoddy lined up for News 24 feedback show|url=
    ↪ https://www.theguardian.com/media/2004/aug/10/bbc.pressandpublishing|
    ↪ newspaper=[[The Guardian]]|publisher=TheGuardian.com / Guardian News
    ↪ and Media Limited|accessdate=31 August 2016}}</ref><ref name=
    ↪ NewsWatch>{{cite web|editor=Matt Holder|date=4 November 2004|title=
    ↪ Welcome to NewsWatch|url=http://news.BBC.co.uk/newswatch/ukfs/hi/
    ↪ newsid_3950000/newsid_3959100/3959141.stm|website=news.BBC.co.uk|
    ↪ publisher=[[BBC News]] / [[BBC Online]]|accessdate=28 August 2016}}</
    ↪ ref> Snoddy started his journalistic career writing for a number of
    ↪ publications on issues relating to the [[news]] industry, and
    ↪ continues in this vein.

==Life and career==
Born in [[Larne]], [[County Antrim]], Northern Ireland, Snoddy was educated
    ↪ at [[Larne Grammar School]], and [[Queen's University Belfast|Queen's
    ↪  University]] in [[Belfast]].<ref name=BBC_Snoddy/> After university,
    ↪  he worked on local and regional [[newspaper]]s, before joining ''[[
    ↪ The Times]]'' in 1971.<ref name=BBC_Snoddy/> He later moved to the
    ↪ ''[[Financial Times]]'' (FT),<ref name=Guardian/> joining in 1978,
    ↪ and reporting on media issues for the paper, before returning to ''
    ↪ The Times'' as media editor in 1997.<ref name=BBC_Snoddy/><ref name=
    ↪ Guardian/> Whilst working at the FT, Snoddy made occasional
    ↪ appearances as [[guest presenter]] on the observational newspaper
    ↪ review TV show ''[[What the Papers Say]]''.<ref>{{cite web|title=What
```

```
    ↪  the Papers Say (TV Series -1956 )|url=http://www.IMDb.com/title/
    ↪ tt0273032/?ref_=nm_knf_t2|website=www.IMDb.com|accessdate=30 August
    ↪ 2016}}</ref> At present, Snoddy is a [[freelance journalist]],
    ↪ writing predominantly for ''[[The Independent]]'',<ref name=Guardian
    ↪ /> although his articles sometimes appear in other newspapers and
    ↪ publications.<ref name=BBC_Snoddy/>
48
49  Following his departure from ''The Times'' in late June 2004,<ref name=
    ↪ Guardian/> Snoddy presented ''[[NewsWatch (TV series)|NewsWatch]]''
    ↪ from its inception in 2004 to 2013.  The programme, now titled as ''
    ↪ Newswatch'',<ref>{{cite web|url=http://www.BBC.co.uk/programmes/
    ↪ b00qjrk2|title=BBC News Channel - Newswatch|website=www.BBC.co.uk|
    ↪ publisher=[[BBC News]]|accessdate=28 August 2016}}</ref> was launched
    ↪  as a response to the [[Hutton Inquiry]], as part of an initiative to
    ↪  make [[BBC News]] more accountable.<ref name=NewsWatch/> His other
    ↪ television work has included presenting [[Channel 4]]'s award-winning
    ↪  series ''Hard News'', which covered the press, and [[Sky News]]' ''
    ↪ Media Monthly''.<ref name=BBC_Snoddy/>
50
51  In addition, Snoddy is the author of a [[biography]] of the media tycoon [[
    ↪ Michael Green (television magnate)|Michael Green]]: ''The Good, the
    ↪ Bad and the Unacceptable: The Hard News about the British Press'',
    ↪ about ethics in the newspaper industry,<ref name=BBC_Snoddy/><ref
    ↪ name=GBU>{{cite book|author=Raymond Snoddy|title=The Good, the Bad
    ↪ and the Unacceptable: The Hard News about the British Press|url=https
    ↪ ://books.google.com/books?id=W1xoQgAACAAJ|year=1992|publisher=[[Faber
    ↪  and Faber|Faber & Faber]]|isbn=9780571161539|accessdate=28 August
    ↪ 2016}}</ref> and other books.<ref name=GoogleUKbooksearch>{{cite web|
    ↪ title=inauthor:"Raymond Snoddy" - Google UK book search|url=https://
    ↪ www.Google.co.uk/search?tbo=p&tbm=bks&q=inauthor:%22Raymond+Snoddy
    ↪ %22|website=www.Google.co.uk|accessdate=28 August 2016}}</ref>
52
53  Whilst Media Editor at ''The Times'' in 2000, Snoddy was awarded the honour
    ↪ of the [[Order of the British Empire|Ordinary Officer of the Civil
    ↪ Division of the said Most Excellent Order of the British Empire]] (
    ↪ OBE); for his services to journalism.<ref name=LondonGazette/>
54
55  ==Bibliography==
56  *1993: ''The Good, the Bad and the Unacceptable: the hard news about the
    ↪ British press'', Faber & Faber, {{ISBN|978-0571161539}}<ref name=
    ↪ GoogleUKbooksearch/>
57  *1996: ''Greenfinger: the rise of Michael Green and Carlton Communications
    ↪ '', Faber & Faber, {{ISBN|978-0571173747}}<ref name=
    ↪ GoogleUKbooksearch/>
58  *2001: ''It Could Be You: the untold story of the National Lottery'', Faber
    ↪ & Faber, {{ISBN|978-0571200870}}<ref name=GoogleUKbooksearch/>
59
60  ==References==
61  {{reflist}}
62
63  {{s-start}}
64  {{s-media}}
65  {{s-bef|before=new position}}
66  {{s-ttl|title=sole presenter of [[BBC]] [[BBC News (TV channel)|News 24]]
    ↪ ''[[NewsWatch (TV series)|NewsWatch]]''|years=2004 - 2012}}
```

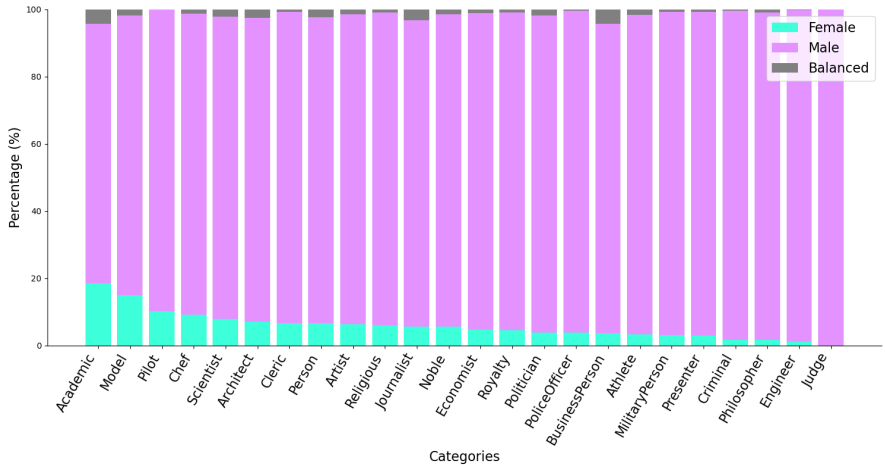
```
{{s-aft|after=[[Samira Ahmed]]}}
{{s-end}}

{{BBC News}}

{{Authority control}}

{{DEFAULTSORT:Snoddy, Raymond}}
[[Category:1946 births]]
[[Category:People from Larne]]
[[Category:People educated at Larne Grammar School]]
[[Category:Alumni of Queen's University Belfast]]
[[Category:British male journalists]]
[[Category:Journalists from Northern Ireland]]
[[Category:BBC newsreaders and journalists]]
[[Category:Television presenters from Northern Ireland]]
[[Category:Officers of the Order of the British Empire]]
[[Category:Living people]]
[[Category:Columnists from Northern Ireland]]
[[Category:Male non-fiction writers from Northern Ireland]]


{{UK-journalist-stub}}
```

Listing A.1: Wikitext example of a specific page and revision.

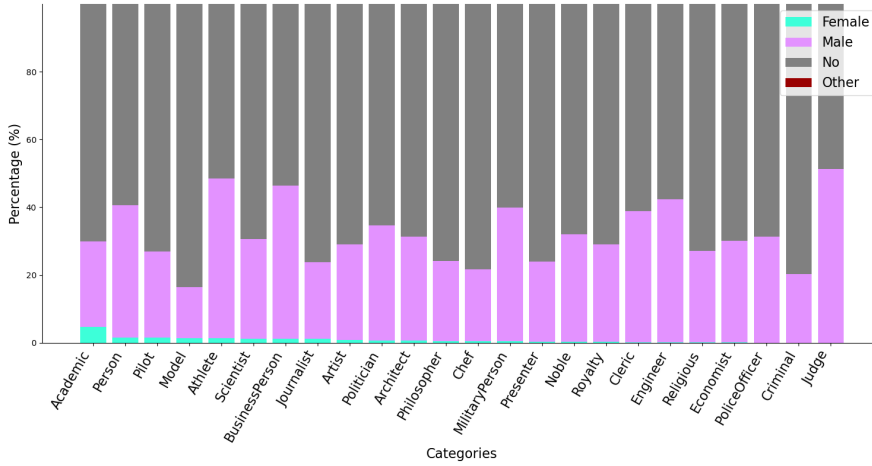
APPENDIX B

Data description

B.1 Gender and category distribution



((a)) Gender of the group measured as gender majority.



((b)) Gender of the group measured as gender totality. "No" label designate when an article has a mixture of editors' genders.

Figure B.1: Distribution of group of editors' gender across different categories. The categories are sorted in descendent order of percentage of female.

APPENDIX C

Results

C.1 Classification of texts

Top positive features		Top negative features	
woman	8.736	john	-2.329
husband	3.483	football	-2.281
mary	3.149	footballer	-2.105
life	2.999	ndash	-2.030
early	2.806	charles	-1.752
daughter	2.495	chairman	-1.729
career	2.381	men	-1.689
personal	2.329	game	-1.578
education	2.302	son	-1.528
riksdag	2.229	william	-1.427

((a)) Class: Female page and Female editors

Top positive features		Top negative features	
woman	6.458	footballer	-1.482
actress	5.548	david	-1.511
female	3.435	irish	-1.598
mother	2.985	james	-1.598
husband	2.931	painter	-1.642
sister	2.343	son	-1.861
child	2.343	actor	-2.010
girl	1.892	william	-2.094
lady	1.825	john	-2.578
opera	1.770	died	-2.648

((b)) Class: Female page and Male editors .

Top positive features		Top negative features	
refer	3.054	york	-1.713
john	2.938	list	-1.713
election	2.697	career	-1.745
james	2.410	life	-1.772
painter	2.237	mother	-1.780
rkd	2.121	first	-2.051
sergey	1.962	female	-2.264
ndash	1.956	husband	-2.438
william	1.934	actress	-3.163
parliament	1.879	woman	-6.684

((c)) Class: Male page and Female editors

Top positive features		Top negative features	
son	2.879	sister	-2.169
footballer	2.294	reference	-2.211
charles	2.294	mary	-2.366
football	2.230	election	-2.384
chairman	2.117	mother	-2.547
peter	2.078	daughter	-2.618
john	1.969	female	-2.911
bishop	1.930	husband	-3.976
norwegian	1.823	actress	-4.061
director	1.791	woman	-8.509

((d)) Class: Males page and Male editors .

Figure C.1: Top positive and negative features with their coefficients for the logistic regression classification of four classes.

C.2 Number of words

Category	Female editors	Male editors
Person	(268.447, 276.957)	(274.057, 284.277)
Athlete	(57.169, 61.449)	(106.401, 118.053)
Politician	(215.157, 238.547)	(210.749, 235.906)
Artist	(369.372, 399.528)	(370.025, 406.158)

((a)) Female biographies.

Category	Female editors	Male editors
Person	(72.0108, 77.027)	(256.627, 267.833)
Athlete	(45.270, 48.786)	(165.448, 178.314)
Politician	(189.822, 211.575)	(250.918, 275.340)
Artist	(156.505, 182.286)	(452.008, 534.959)

((b)) Male biographies.

Figure C.2: Confidence intervals for the median of the word count distribution across the four largest categories.

C.3 Tf-idf scores

Term	Tf-idf score
university	0.4753
research	0.3387
health	0.1917
professor	0.1856
american	0.1715
education	0.1305
school	0.1271
women	0.1234
award	0.1188
degree	0.1155

((a)) Female biographies by female editors.

Term	Tf-idf score
university	0.5547
research	0.2513
professor	0.1919
american	0.1549
studies	0.1360
education	0.1196
work	0.1172
new	0.1117
history	0.1103
college	0.1098

((b)) Female biographies by male editors.

Term	Tf-idf score
university	0.4525
research	0.2889
professor	0.1914
dugan	0.1725
lowary	0.1568
engineering	0.1392
ronen	0.1333
american	0.1288
school	0.1218
degree	0.1183

((c)) Male biographies by female editors.

Term	Tf-idf score
university	0.5109
research	0.1835
professor	0.1794
history	0.1742
new	0.1622
american	0.1401
studies	0.1273
press	0.1046
school	0.1009
college	0.0926

((d)) Male biographies by male editors.

Figure C.3: 10 highest tf-idf scores with their correspondent words for each class in the category *Academic*.

Term	Tf-idf score
women	0.2605
university	0.2398
born	0.2028
life	0.1723
first	0.1639
new	0.1609
school	0.1506
de	0.1441
american	0.1424
work	0.1388

((a)) Female biographies by female editors.

Term	Tf-idf score
align	0.2063
style	0.1898
women	0.1741
born	0.1725
university	0.1678
film	0.1541
first	0.1495
new	0.1416
de	0.1405
center	0.1328

((b)) Female biographies by male editors.

Term	Tf-idf score
born	0.2388
de	0.2201
university	0.1895
redirect	0.1379
member	0.1312
first	0.1224
became	0.1121
died	0.1102
new	0.1089
election	0.1072

((c)) Male biographies by female editors.

Term	Tf-idf score
de	0.1961
university	0.1915
new	0.1907
born	0.1789
first	0.1609
one	0.1231
life	0.1100
became	0.1072
years	0.1057
american	0.1005

((d)) Male biographies by male editors.

Figure C.4: 10 highest tf-idf scores with their correspondent words for each class in the category *Person*.

C.4 Bigrams of words

Bigram	Frequency	Bigram	Frequency
('new', 'york')	1555	('early', 'life')	1400
('personal', 'life')	911	('align', 'center')	886
('united', 'states')	848	('life', 'education')	632
('high', 'school')	551	('world', 'war')	478
('de', 'la')	471	('scope', 'col')	419
('class', 'wikitable')	406	('scope', 'row')	383
('rowspan', '2')	373	('york', 'city')	367
('first', '**woman**')	363	('text', 'align')	335
('align', 'left')	335	('**women**', 'suffrage')	322
('master', 'degree')	297	('style', 'text')	295
('san', 'francisco')	293	('style', 'background')	289
('vice', 'president')	263	('los', 'angeles')	250
('new', 'zealand')	246	('state', 'university')	246
('summer', 'olympics')	234	('selected', 'works')	221
('african', 'american')	220	('two', 'years')	210
('bachelor', 'degree')	207	('half', 'marathon')	207
('war', 'ii')	203	('best', 'known')	196
('became', 'first')	195	('years', 'old')	194
('w', 'c')	187	('fine', 'arts')	186
('university', 'press')	185	('c', 'u')	183
('new', 'jersey')	180	('**woman**', 'suffrage')	171
('one', 'first')	171	('first', '**female**')	168
('civil', 'war')	167	('washington', 'c')	166
('human', 'rights')	166	('**women**', 'rights')	164
('hall', 'fame')	163	('years', 'later')	162

Table C.1: Bigram Frequencies for Fpage Feditors in Person category

Bigram	Frequency	Bigram	Frequency
('align', 'center')	30542	('align', 'left')	23717
('text', 'align')	21668	('style', 'background')	20480
('style', 'text')	19842	('new', 'york')	18479
('class', 'wikitable')	15644	('rowspan', '2')	14337
('data', 'sort')	11598	('sort', 'value')	11485
('value', '99')	10862	('99', 'data')	9369
('united', 'states')	8278	('early', 'life')	8079
('scope', 'row')	7765	('scope', 'col')	7600

Bigram	Frequency	Bigram	Frequency
('0', '0')	7579	('style', 'width')	6105
('personal', 'life')	5998	('high', 'school')	5892
('tv', 'series')	5594	('title', 'role')	5199
('year', 'title')	5159	('wikitable', 'sortable')	4895
('los', 'angeles')	4479	('de', 'la')	4447
('film', 'festival')	4439	('york', 'city')	4245
('wikitable', 'style')	3693	('rowspan', '3')	3599
('class', 'unsortable')	3514	('1', '1')	3352
('award', 'best')	3258	('wikitable', 'year')	3246
('1', '0')	3220	('world', 'war')	3176
('best', 'known')	3168	('0', '1')	3042
('valign', 'top')	3017	('background', 'fff')	2957
('short', 'film')	2938	('align', 'right')	2904
('new', 'zealand')	2875	('best', 'actress')	2697
('world', 'cup')	2680	('television', 'series')	2624
('6', '4')	2600	('life', 'education')	2597
('sortable', 'year')	2593	('font', 'size')	2591

Table C.2: Bigram Frequencies for Fpage Meditors in Person category

Bigram	Frequency	Bigram	Frequency
('summer', 'olympics')	420	('align', 'center')	392
('general', 'election')	308	('may', 'refer')	296
('new', 'york')	258	('de', 'la')	200
('early', 'life')	198	('united', 'states')	178
('world', 'war')	175	('scope', 'row')	171
('rowspan', '2')	158	('member', 'parliament')	140
('work', 'part')	140	('legislative', 'assembly')	138
('art', 'competition')	131	('event', 'art')	130
('class', 'wikitable')	130	('rkd', 'nl')	126
('dispatcher', 'aspx')	122	('aspx', 'actionsearch')	122
('nl', 'rkddb')	121	('rkddb', 'dispatcher')	121
('actionsearch', 'databasechoiceartists')	120	('golden', 'age')	119
('dutch', 'golden')	117	('tamil', 'nadu')	116
('sports', 'reference')	116	('de', 'l')	114
('university', 'press')	113	('half', 'marathon')	113
('according', 'rkd')	111	('iaaf', 'org')	110

Bigram	Frequency	Bigram	Frequency
('van', 'der')	110	('los', 'angeles')	108
('hong', 'kong')	107	('player', 'competed')	107
('painter', 'biography')	105	('people', 'deputy')	102
('labour', 'party')	100	('high', 'school')	93
('northern', 'ireland')	92	('personal', 'life')	91
('scope', 'col')	90	('dbnl', 'org')	89
('biography', 'according')	87	('org', 'tekst')	87
('best', 'known')	87	('servant', 'people')	86
('digital', 'library')	85	('iaaf', 'retrieved')	84

Table C.3: Bigram Frequencies for Mpage Feditors in Person category

Bigram	Frequency	Bigram	Frequency
('new', 'york')	61637	('united', 'states')	28871
('style', 'background')	27088	('align', 'center')	19485
('world', 'war')	19172	('de', 'la')	17611
('early', 'life')	15912	('align', 'left')	15780
('high', 'school')	13080	('rowspan', '2')	11472
('text', 'align')	11285	('york', 'city')	10969
('style', 'text')	10776	('los', 'angeles')	10730
('class', 'wikitable')	10659	('university', 'press')	10274
('war', 'ii')	10127	('best', 'known')	9205
('pp', 'nbsp')	8749	('personal', 'life')	8442
('two', 'years')	7542	('new', 'zealand')	7509
('de', 'l')	7383	('film', 'festival')	7279
('0', '0')	6833	('vice', 'president')	6794
('years', 'later')	6721	('state', 'university')	6406
('p', 'nbsp')	6368	('co', 'uk')	6137
('york', 'times')	6039	('new', 'jersey')	5859
('san', 'francisco')	5812	('tv', 'series')	5701
('jpg', 'thumb')	5488	('scope', 'row')	5345
('background', 'cfcfff')	5263	('background', 'dfffdf')	5230
('well', 'known')	5213	('summer', 'olympics')	5117
('civil', 'war')	5057	('award', 'best')	4996
('hall', 'fame')	4966	('first', 'class')	4965
('following', 'year')	4802	('fine', 'arts')	4746
('three', 'years')	4609	('background', 'efcfff')	4559
('index', 'php')	4439	('supreme', 'court')	4414

Table C.4: Bigram Frequencies for Mpage Meditors in Person category

C.5 Hyperlinks structure

ID	**Category**	**Female pages**	**Male pages**	**p-value**
0	Academic	0.725901	0.527119	<0.001 ***
1	Architect	0.541667	0.415878	<0.001 ***
2	Artist	0.303882	0.210164	<0.001 ***
3	Athlete	0.275541	0.276346	0.7
4	BusinessPerson	0.987952	0.954023	0.113
5	Chef	0.405063	0.320896	0.03 *
6	Cleric	0.166667	0.098232	<0.001 ***
7	Criminal	0.608911	0.439362	<0.001 ***
8	Economist	0.458101	0.244382	<0.001 ***
9	Engineer	0.809524	0.493488	0.006 **
10	Journalist	0.893333	0.868516	0.217
11	Judge	0.916667	0.920000	1.0
12	MilitaryPerson	0.445732	0.301772	<0.001 ***
13	Model	0.509128	0.580153	0.12
14	Noble	0.287026	0.272727	0.448
15	Person	0.310058	0.296524	<0.001 ***
16	Philosopher	0.265517	0.149633	<0.001 ***
17	Pilot	0.703125	0.556391	0.059
18	PoliceOfficer	0.600000	0.606061	0.827
19	Politician	0.146320	0.100529	<0.001 ***
20	Presenter	0.709677	0.643646	0.141
21	Religious	0.311178	0.164590	<0.001 ***
22	Royalty	0.030661	0.041450	0.005 **
23	Scientist	0.490076	0.276225	<0.001 ***

Table C.5: Isolation ratio for the pages in every category. The last column contains the p-values. Statistical comparisons were performed using a permutation test where null hypothesis is that the isolation ratio is the same under both genders of pages (*** $p < 0.001$, ** $p < 0.01$, * $p<0.05$).

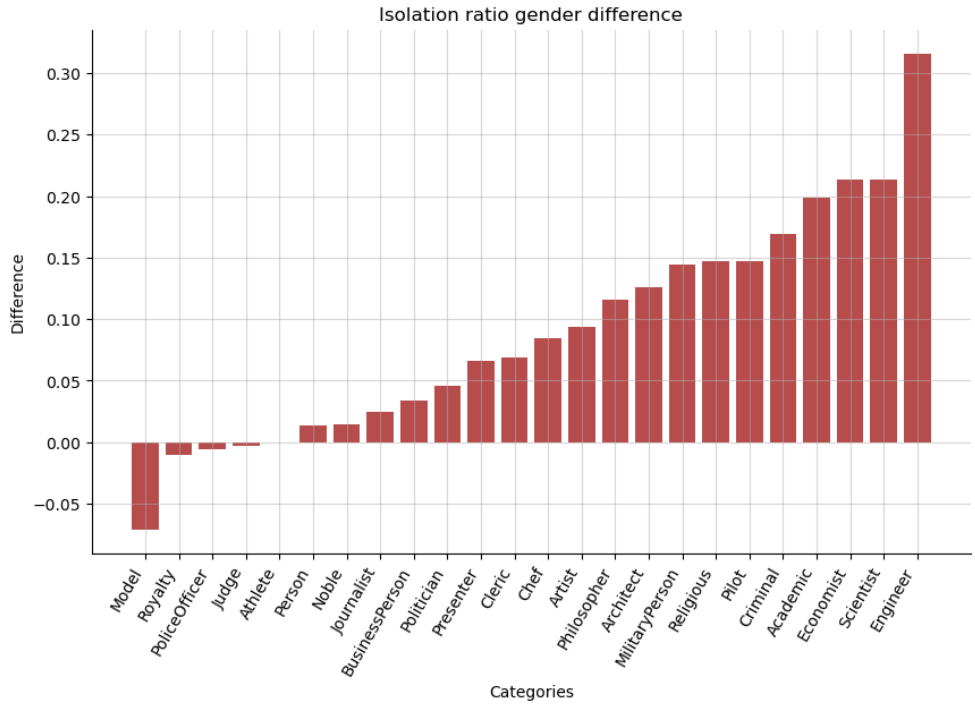


Figure C.5: Difference in isolation ratio ($I_F - I_M$) sorted in ascending order per category.